POPULATION GENOMICS, HYBRID STERILITY, & GENE EXPRESSION IN THE *ANOPHELES GAMBIAE* SPECIES COMPLEX

A Dissertation

by

KEVIN CANNING DEITZ

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Michel A. Slotman
Committee Members,	J. Spencer Johnston
	Craig J. Coates
	William B. Murphy
Head of Department,	David W. Ragsdale

August 2017

Major Subject: Entomology

ABSTRACT

This dissertation explores the genomics of speciation in the *An. gambiae* complex. The complex is comprised of eight recently diverged species and contains the most important vectors of human malaria. Male F1 hybrids between these species are completely sterile and females exhibit varying levels of sterility. I performed a population genomic analysis of the three genetic clusters that comprise *An. melas* that illuminated genome-wide divergence between them. This is suggestive of a pattern of allopatric divergence with little to no gene flow.

I analyzed gene expression data to assess if dosage compensation occurs in males of *An. coluzzii, An. arabiensis*, and *An. quadriannulatus*, in order to balance expression between the hemizygous X chromosome and the autosomes. Dosage compensation is acting in each species and is not influenced by hybridization. I identified male- and female-biased genes in each parental species and mis-expressed genes in F1 hybrid males that are involved in hybrid sterility. Mis-expressed genes are involved in mitosis, spermatogenesis, and other physiological processes involved in the sterility phenotype. An analysis of allele-specific expression found that *An. coluzzii* and *An. arabiensis* have a high proportion of genes (~80%) that are conserved or compensatory in their transcription. In contrast, *An. coluzzii* and *An. quadriannulatus* have a higher proportion of genes that have diverged in trans-. These patterns may reflect higher levels of historical and ongoing introgression between *An. coluzzii* and *An. arabiensis* than between *An. coluzzi* and *An. quadriannulatus.*

Lastly, I performed a QTL analysis of sterility in a (*An. coluzzi* x *An. quadriannulatus*) x *An. quadriannulatus* backcross to identify regions of the *An. coluzzi* genome that are responsible for male sterility when introgressed into an *An. quadriannulatus* genomic background. Five autosomal QTL were identified that interact through epistatic interactions with each other and the X chromosome, which is responsible for the majority of the variation observed in the sterility phenotype. Two autosomal QTL of large effect are shared with the *An. coluzzi* x *An. arabiensis* cross,

indicating that these regions of the *An. coluzzii* genome contribute to reproductive isolation between this species and multiple members of the *An. gambiae* species complex.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Michel A. Slotman, and my committee members, Dr. J. Spencer Johnston, Dr. Craig J. Coates, and Dr. William J. Murphy, for their guidance and support throughout the course of this research.

CONTRIBUTORS AND FUNDING SOURCES

This work was supervised by a dissertation committee consisting of Dr. Michel A. Slotman, Dr. J. Spencer Johnston, and Dr. Craig J. Coates of the Department of Entomology and Dr. William J. Murphy of the Department of Veterinary and Integrative Biosciences.

Mosquitoes sampled for Chapter I were collected by the staff of the Bioko Island Bioko Island Malaria Control Project (BIMCP). Collections on Bioko Island were conducted as part of the vector monitoring efforts under the BIMCP. Collaborators Parfait H. Awono-Ambene, Christophe Antonio-Nkondjo, and Frederic Simard sampled mosquitoes in Ipono, Cameroon. DNA sequencing for Chapter I was performed by the *Anopheles* Genome Consortium, led by Michael C. Fontaine of the University of Groningen, Daniel E. Neafsey of the Broad Institute of MIT and Harvard, and Nora J. Besansky of the University of Notre Dame. All other work conducted for the dissertation was completed by the student independently.

DNA sequencing and genotyping of backcross mosquitoes used in the QTL analysis of sterility loci was funded by a National Science Foundation Doctoral Dissertation Improvement Grant (award #1601675) awarded to Michel A. Slotman and Kevin C. Deitz and a Texas EcoLab grant awarded to Kevin C. Deitz. RNA sequencing of F1 hybrid and parental strain mosquitoes was funded by a Texas A&M University Genomics Seed Grant awarded to Michel A. Slotman, Kevin C. Deitz, and Luciano V. Cosme.

This work was partially supported by teaching assistantships through the Department of Entomology, the J.H. Benedict, Sr. Memorial Graduate Student Scholarship, the Herb Dean '40 Endowed Scholarship, and the Office of Graduate and Professional Studies Dissertation Fellowship.

TABLE OF CONTENTS

Page

CHAPTER I

GENOME-WIDE DIVERGENCE IN THE WEST-AFRICAN MALARIA VECTOR *ANOPHELES MELAS**

Introduction

The *Anopheles gambiae* complex of African malaria mosquitoes is a model system for the study of speciation (Fontaine *et al.,* 2015, Mallet *et al.,* 2015, Neafsey *et al.,* 2015, Nosil, 2012). This is partly due to its importance to human health, but also because varying levels of reproductive isolation and introgression are found between its member species (Besansky *et al.,* 1994, Davidson, 1962, Fontaine *et al.,* 2015, Lanzaro and Lee, 2013, Marsden *et al.,* 2011, Powell *et al.,* 1999, Slotman *et al.,* 2004, 2005a, 2005b, Weetman *et al.,* 2014), chromosomal and molecular forms occur within species (Coluzzi *et al.,* 2002, della Torre *et al.,* 2001, Favia *et al.,* 2001, Gentile *et al.,* 2001, White *et al.,* 2011), and contrasting patterns of intraspecific population structure have been observed between species (Deitz *et al.,* 2012, Donnelly and Townson, 2000, Lehman *et al.,* 2003, Loaiza *et al.,* 2012). The recent evolutionary analyses of 16 *Anopheles* genomes highlighted the role of adaptive introgression in the divergence of the *An. gambiae* complex (Clarkson *et al.*, 2014, Fontaine *et al.*, 2015, Norris *et al.*, 2015), and how biological factors involved in their capacity to vector human malaria parasites have influenced the evolution of these species (Neafsey *et al.,* 2015).

Eight species have now been formerly described within the *An. gambiae* complex, including two recent additions: *An. coluzzii,* formerly *An. gambiae* M molecular form, and *An. amharicus,* formerly *An. quadriannulatus* B (Coetzee *et al.,* 2013). The elevation of the *An. gambiae* M form to species rank was based on ecological divergence, assortative mating (della Torre *et al.,* 2001, Simard *et al.,* 2009, Tripet *et al.,* 2005, Aboagye-Antwi *et al.,* 2015), and genetic divergence that appears to be limited to several small regions of the genome (Turner *et al.,* 2005, White *et al.*, 2010). The

*Reprinted with permission from Deitz KC, Athrey GA, Jawara M, Overgaard HJ, Matias A, Slotman MA (2016) Genome-Wide Divergence in the West-African Malaria Vector Anopheles melas. *G3: Genes, Genomes, Genetics* **8**: 2867-2879.

description of *An. coluzzii* therefore broke with the tradition of describing new species in the complex based on the presence of hybrid sterility (Davidson, 1964, Hunt *et al.,* 1998), as hybrids between *An. gambiae* and *An. coluzzii* are fully fertile (Diabaté *et al.,* 2007). Thus, the description of *An. coluzzii* is aligned more with a genotypic cluster species concept (Mallet 1995) rather than a biological species concept (Mayr 1970).

A recent study on the population structure of *An. melas* throughout its range uncovered species-level genetic divergence between three population clusters (Deitz *et al.,* 2012). *An. melas* is distributed along the west coast of Africa as its larval ecology is tied to brackish water mangrove forests and salt marshes. Nonetheless, it is an important vector of human malaria where it is found (Bryan *et al.,* 1987, Caputo *et al.,* 2008), with the average number of malaria infective *An. melas* bites/person/year sometimes reaching 130 (Overgaard *et al.,* 2012). Coluzzi *et al.* (2002) found that some chromosomal inversions were non-randomly distributed between *An. melas* populations, suggesting the presence of some reproductive barriers. Deitz *et al.* (2012) showed that *An. melas* is in fact divided into three genetic clusters that appear to be mostly isolated from each other. Two of these clusters are distributed on the African mainland: *An. melas* West ranges from The Gambia to Northwest Cameroon, and *An. melas* South ranges from Southeast Cameroon to Angola. A third cluster, *An. melas* Bioko, is limited to Bioko Island, Equatorial Guinea, located approximately 40 km off the Cameroonian coast (Figure 1).

No mtDNA haplotypes are shared between *An. melas* clusters, and microsatellite data indicates almost complete genetic isolation, with the exception of limited introgression into *An. melas* West from the South and Bioko, which was identified through a Bayesian analysis of population structure. Additionally, the level of genetic divergence (F_{ST}) between *An. melas* West and South equaled or exceeded levels previously observed between *An. gambiae* and *An. arabiensis* (Slotman *et al.,* 2005a, Fontaine *et al.,* 2015). Interestingly, *An. melas* West and South populations are only separated by approximately 190 km of unsampled terrain along the Cameroonian coast. The high level of isolation of the *An. melas* Bioko Island population is also remarkable given the short distance to the mainland, and the very low level of genetic differentiation

between Bioko Island and mainland populations of both *An. gambiae* and *An. coluzzii* (Moreno *et al.,* 2007, Deitz *et al.,* 2012).

An analysis of the demographic history of *An. melas* populations using approximate Bayesian computation analysis indicated that a larger ancestral *An. melas* population split into two mainland clusters through a vicariance event sometime during the last several hundred thousand years. Similarly, *An. melas* Bioko was once connected to *An. melas* West populations, but became isolated around 90,000 years before present, presumably due to rising sea levels (Deitz *et al.,* 2012).

In the present study we used a whole-genome, pooled-population sequencing (Pool-seq) approach (Schlötterer *et al.,* 2014) to examine genome-wide patterns of diversity within, and divergence between, a representative population sample of *An. melas* West, South, and Bioko. Such an analysis may reveal whether the geographically isolated forms of *An. melas* harbor any genetically highly diverged regions of the genomes, similar to those that have been tied to pre-mating isolation between *An. gambiae s.s.* and *An. coluzzii* (Aboagye-Antwi *et al*., 2015). The genome-wide SNP data show that *An. melas* population clusters have high levels of genome-wide genetic differentiation, as evidenced by numerous high-F_{ST} and fixed SNPs in each population comparison. Genetic differentiation is particularly high on the X chromosome, which also carries the largest number of fixed differences. Additionally, we identified candidate regions under positive selection within each *An. melas* population cluster. A lack of narrow, highly differentiated genomic regions is consistent with allopatric divergence with little or no introgression.

Methods

Population Genomic Analysis

Pool-seq was performed on DNA of *Anopheles melas* females collected from Ballingho, The Gambia (*N*=20), Ipono, Cameroon (*N*=23), and Arena Blanca, Bioko Island, Equatorial Guinea (*N*=20). These populations fall within *An. melas* West, South and Bioko Island genetic clusters, respectively (Figure 1) (Deitz *et al.,* 2012). Populations were chosen based upon the high quality of DNA available to create pooled

libraries for sequencing, and the lack of gene flow observed between them and neighboring *An. melas* clusters (Deitz *et al.*, 2012). Mosquito collection and DNA extraction methods are as described in Deitz *et al.* (2012). We pooled equal amounts of DNA from each individual, and sequencing libraries were constructed from 1.0 μg of pooled DNA. Covaris shearing (Fisher *et al.,* 2011) was used to produce approximately 200 bp inserts for each library. Libraries were bar-coded, combined, and paired-end sequenced on a single lane of the Illumina HiSeq 2000 DNA sequencing platform.

Sequencing reads were trimmed to a minimum Phred quality score of 20 and a minimum length of 50 base pairs using Trimmomatic version 0.35 (Bolger *et al.,* 2014), and then mapped to the *An. gambiae* PEST P4.3 genome assembly (Holt *et al.,* 2002) using Stampy (Lunter and Goodson, 2011) with a substitution rate = 0.02. Stampy is designed to map DNA sequencing reads to a divergent reference genome and has been previously used for this purpose in the *An. gambiae* species complex (Smith *et al.,* 2015). Sequencing reads were mapped to the *An. gambiae* genome rather than the *An. melas* genome (Neafsey *et al.*, 2015) because the former is assembled into chromosomes and at the present time the *An. melas* genome is comprised of 20,229 scaffolds (Giraldo-Calderón *et al.,* 2015, Neafsey *et al.,* 2015). No coordinate lift-over file is available to convert the coordinates of the *An. melas* scaffolds to those of the *An. gamibae* P4.3 chromosomes. As such, we aligned our data to the *An. gambiae* genome because it allowed us to interpret population genetic statistics in the context of chromosomal location. SAM alignment files were sorted, converted to BAM format, filtered to a minimum mapping quality value (MAPQ) of 20, and converted to pileup files using SAMtools version 0.1.19 (Li *et al.,* 2002).

Pileup files were used to calculate nucleotide diversity (π*,* Nei and Li, 1979) and Tajima's *D* (Tajima, 1989) using the PoPoolation package (Kofler *et al.,* 2011a). Both statistics were calculated using 100 kb, non-overlapping, sliding-windows using a minimum sequence coverage of four reads and maximum coverage of 40. We required a minimum of two reads for each allele at a polymorphic site to retain the site for further analysis. The highly repetitive nature of heterochromatic genomic regions leads to

inaccurate read mapping, which biases population genetic statistics. Heterochromatic regions of the *An. gambiae* reference genome (Sharakhova *et al.,* 2010) were removed for the calculation of π, Tajima's *D*, and F_{ST} summary statistics. Vertical grey bars in Figures 3 and 4 highlight heterochromatic regions.

Multiple pileup files were created with SAMtools version 0.1.19 (Li *et al.,* 2002) and transformed into synchronized pileup files using Popoolation2 (Kofler *et al.,* 2011b). This program was then used to calculate pair-wise F_{ST} values for each single nucleotide polymorphism (SNP), and for 100 kb, non-overlapping sliding-windows using a minimum sequencing depth of 30x and a maximum equal to the top 2% of the sequencing depth distribution of each pool. Reads exceeding the top 2% sequencing depth threshold were excluded from our analysis to reduce the effect of sequencing and mapping bias.

We chose 30x coverage to measure SNP and window-based F_{ST} because it allows us to have enough coverage in both populations in a comparison to provide a genome-wide distribution of informative loci for population genomic analysis, and have enough the power to detect significant differentiation. In our initial F_{ST} null distribution simulations we found that coverage below this value incorporates a high level of variation in the allele frequency and F_{ST} estimates at a single locus. Thus, a high coverage threshold allows us to be confident that differences in read coverage between populations in a comparison is not biasing our F_{ST} calculation. We used a lower threshold for π and Tajima's *D* (above) because these values are averaged over a 100kb window and inaccuracy in estimates for individual loci should cancel out within each window and not introduce bias.

If significant SNPs fell within the bottom 5% of the Tajima's *D* distribution in both populations in a pair-wise comparison (e.g., *An. melas* West and South), the SNP was subjected to gene ontology analyses. These analyses excluded SNPs and low Tajima's *D* regions that fell inside regions of heterochromatin in the *An. gambiae* reference genome. SNPs were compared to the *An. gambiae,* AgamP4.4 gene set (Holt *et al.,* 2002, Sharakhova *et al.,* 2007) to determine if they fell within a known gene exon.

The molecular function, biological process, and protein class of these genes was determined using the Panther Classification System (Thomas *et al.*, 2003, Mi *et al.*, 2010).

To identify regions of introgression between *An. melas* forms, we calculated Patterson's *D*-statistic, i.e. the ABBA/BABA test (Green *et al.,* 2010, Durand *et al.,* 2011), using the program ANGSD (Korneliussen *et al.,* 2014). We used 100-kb windows to analyze patterns of introgression between *An. melas* populations throughout the genome. The ABBA/BABA test compares biased proportions of ABBA vs. BABA patterns across a four species lineage to identify regions of introgression between populations P_3 and P_1 or P_3 and P_2, given the following topology: $((P_1, P_2)P_3)O)$, where O signifies the outgroup. Positive values of Patterson's *D*-statistic indicate biased proportions of ABBA patterns, indicating introgression between P_3 and P_2, whereas negative Patterson's *D*-statistic values indicate a biased proportion of BABA patterns, and introgression between species P_3 and P_1. It is important to note that this test cannot determine the direction of introgression (i.e, from P_3 to P_1, or P_1 to P_3).

Patterson's *D*-statistic was calculated using *An. gambiae* as an outgroup and using the following tree topology: ((West, Bioko) South) *An. gambiae*). This tree topology is strongly supported by an approximate Bayesian computation analysis of the demographic history of *An. melas* populations based upon microsatellite data (posterior probability = 0.97) (Deitz *et al.,* 2012). This tree topology allowed us to test which scenario is more likely, introgression between *An. melas* South and Bioko (P_3 and P_2) or between *An. melas* South and West (P_3 and P_1). ABBA/BABA sites were included in this analysis if sequence reads had a minimum map quality score of 30, and the SNP had a minimum base quality score of 30. The ANGSD implementation of the ABBA/BABA test uses one allele sampled from each population. While this could result in a loss of power when implemented using Pool-seq data, it will not bias the number of ABBA vs. BABA sites (R. Nielsen, *personal communication*). A delete-m jackknife approach (Busing *et al.,* 1999) was used to determine the standard error of the mean Patterson's *D*-statistic on each chromosome arm, and the entire genome. We calculated a Z-score to

test if ABBA or BABA counts on each chromosome arm differed significantly from the null hypothesis of Patterson's *D*-statistic = 0 (no excess of ABBA or BABA sites), indicating introgression between two of the populations.

Generation of an F_{ST} Null Distribution and False Discovery Rate

Previous studies using Pool-seq identified divergent genomic regions by visually inspecting sliding-window F_{ST} graphs for high peaks (e.g. Karlsen *et al.,* 2013), or considered SNPs to be significant if they were four standard deviations above the mean value of the *Z*-transformed F_{ST} distribution (e.g. Montague *et al.,* 2014). Others considered SNPs to be significantly differentiated between populations if their pair-wise F_{ST} values fell in the top 0.5% of the F_{ST} distribution, and had a Bonferroni-corrected *p*-value lower than 0.05 when subjected to a Fisher's exact test (Kofler *et al.,* 2011b, Fabian *et al.,* 2012). While conservative approaches such as a Bonferroni correction reduce type I error, they may exclude a large number of biologically significant SNPs from downstream analyses (Darum, 2006). Additionally, relying on the Fisher's exact test implemented in PoPoolation2 for detecting significant differences in allele frequencies does not take into account pool size, which can influence allele frequency estimates. Thus, it only works well for studies in which pool size is considerably larger than sequencing coverage and can be ignored. In case of small pool size it will lead to a potentially large number of false positive results.

Therefore, we created a F_{ST} null distribution by simulating F_{ST} values observed between two samples drawn from a single population, given our pool size and sequence coverage. This null distribution allows us to determine which SNPs are significantly differentiated in our data. We created this null distribution by performing simulations in R (https://www.r-project.org). First, we drew 40 alleles (N=20) from a population of 1,000 individuals with a single SNP at an allele frequency of 0.5. We used an initial allele frequency of 0.5 because this value results in the largest variance of the estimated allele frequency. This step was repeated 10 million times to create our "population pool" allele distribution (Figure 2). This step simulates the pooling of individuals. We then drew 30 alleles (the minimum sequencing coverage (30x) used for SNP-wise and

window-based F_{ST} estimation) from our population pool allele distribution. This step was repeated 10 million times to create the "sequencing pool" allele distribution (Figure 2). This step simulates the random generation of sequencing reads from the Pool-seq DNA library. The simulation of these two sampling steps combined provides the distribution of possible allele frequency estimates.

To obtain the F_{ST} null distribution, we drew two allele frequency values from this allele frequency distribution 10 million times and calculated the allele frequency difference between them (Allele Frequency Difference, Figure 2). We calculated the F_{ST} value for each of these pairs using $F_{ST} = (H_T\text{-}H_S)/(H_T)$, where H_T is the total population heterozygosity, and H_S is the sub-population heterozygosity. This process was also repeated 10 million times to create the "pair-wise F_{ST}" distribution. This F_{ST} null distribution was used to find the F_{ST} value for which the false discovery rate (FDR) ≤ 0.05. For each pair-wise population comparison, this was done by finding the threshold F_{ST}-value for which: (*p*-value * Total SNP number)/(significant SNP number) = 0.05. Here, the “p-value” is the proportion of F_{ST} values above the threshold F_{ST}-value in the null distribution, “total SNP number” is the number of SNPs in the population data set, and “significant SNP number” is the number of SNPs in the population data set with an F_{ST}-value above the threshold. In other words, the numerator is the expected number of false positives, and the denominator is the number of significantly differentiated SNPs in the data set.

Results

Sequence Read Quality Control

The sequencing effort resulted in 78,025,712 paired-end reads for *An. melas* West (Ballingho, The Gambia), 52,594,743 for *An. melas* South (Ipono, Cameroon), and 56,776,632 for *An. melas* Bioko (Arena Blanca, Bioko Island, Equatorial Guinea) (Table 4). Paired-end reads were mapped to the genome only if both forward and reverse reads survived quality and length trimming (Phred ≥ 20, length ≥ 50 bp). Mapped reads with MAPQ values greater than 20, and that mapped to chromosomes X, 2, or 3 were retained for further analysis (West = 52.31%, South = 26.16%, and Bioko = 38.38% of original,

raw reads). These reads had a mean length of 98.7 - 99.1 bp for each population (Table 4). However, the mean, genome-wide read coverage per base pair varied between populations (West = 34.44, South = 17.27, and Bioko = 25.41). This factor limited the number of SNPs that met our criteria of 30X coverage for analysis of F_{ST} between population pools.

Nucleotide Diversity & Evolution

While we used lower thresholds (minimum coverage of 4x) for the calculation of nucleotide diversity and Tajima's *D*, our results show that the mean reads per base pair far exceed these values on all chromosome arms in all populations (Table 4). For example, the lowest observed mean reads per base pair (15.63) was on chromosome arm 3L of *An. melas* South. The 4x threshold was used to maximize the number of variable sites within a 100 kb window included in the calculation of nucleotide diversity and Tajima's *D*. On chromosome arm 3L of *An. melas* South, on average 36.34% of a 100 kb window exceeded the minimum coverage threshold.

Genome-wide nucleotide diversity across 100 kb windows was very similar in *An. melas* West from Ballingho, The Gambia (mean $\pi = 0.0052$, std. error = 4.78×10^{-5}), and *An. melas* South from Ipono, Cameroon (mean $\pi = 0.0048$, std. error = 5.31×10^{-5}), but perhaps not unexpectedly, was somewhat lower in *An. melas* Bioko from Arena Blanca, Bioko Island (mean $\pi = 0.0034$, std. error = 5.12×10^{-5}, Table 1). This pattern was consistent across all chromosomes (*An. melas* West $\pi >$ *An. melas* South $\pi >$ *An. melas* Bioko π) (Table 1, Figure 3). In each population, mean chromosomal nucleotide diversity was higher on the 3rd chromosome, and lowest on *2R* or *X* (Table 1, Figure 3). Interestingly, the patterns of nucleotide diversity are remarkably concordant between *An. melas* populations when viewed across their genomes, with the exception of a peak of high nucleotide diversity on chromosome *2L* in *An. melas* Bioko (Figure 3).

Tajima's *D* was calculated to identify genomic regions that may be evolving under positive selection in each population. Mean Tajima's *D* was negative for all populations, indicating a deviation from neutral evolution ($D = 0$) (Table 1, Figure 3). Various low Tajima's *D* regions are shared between all three populations, although some

low Tajima's *D* windows are unique to a single population (Figure 3). While broad patterns of Tajima's *D* for each population are similar across their genomes, the genome-wide mean Tajima's *D* of *An. melas* West is over three times lower than that of *An. melas* South and Bioko (Table 1, Figure 3).

F_{ST} Null Distribution

To determine significance thresholds for genetic differentiation (F_{ST}) between the three *An. melas* populations, the null distribution of allele frequency differences was determined based on our pooling and sequencing coverage using simulations. Next, two values were randomly drawn from this distribution to calculate an F_{ST} value. Each step of the simulation was repeated 10 million times to create each distribution. The first step in this simulation created a population pool with a mean allele frequency of 0.5 and a range of 0.1-0.9 (Table 2, Figure 2). The second step created a sequencing pool distribution with a mean allele frequency of 0.5 and a range of 0.0 to 1.0. The final pair-wise F_{ST} null distribution ranges from 0.0 to 0.875 and has a mean of 0.046 (Table 2, Figure 2). For each *An. melas* pair-wise population comparison the F_{ST} value corresponding to FDR = 0.05 was determined and set as the significance threshold for the SNP-wise F_{ST} analyses. These significance thresholds between the populations are F_{ST} = 0.463 for West-South, F_{ST} = 0.446 for West-Bioko, and F_{ST} = 0.402 for South-Bioko. While these values are high due to relatively small pool sizes and low sequencing coverage, this conservative approach reduces the number of false positive results.

Genetic Differentiation and Introgression

Significant genetic differentiation between the three *An. melas* population clusters extends across the entire genome (Table 5, Table 3), and includes fixed SNPs on all chromosome arms (Table 3, Figure 3). Even though the Ipono, Cameroon and Arena Blanca, Bioko Island populations, which represent *An. melas* South and Bioko, respectively, are geographically close compared to the Ballingho, The Gambia (*An. melas* West), they are most differentiated (Q1=0.018, median F_{ST}=0.033, mean F_{ST}=0.114, Q3=0.091), followed by the West and Bioko (Q1=0.016, median F_{ST} =0.028, mean F_{ST}=0.076, Q3=0.055), and West and South (Q1=0.021, median F_{ST}=0.034, mean

F_{ST}=0.075, Q3=0.062) (Table 5). *An. melas* South and Bioko also have the highest number of significantly differentiated (39,730, 8.56% of total) and fixed SNPs (5,387, 1.16% of total) between them (total SNPs = 463,910), followed by West and Bioko (significant = 21,427 (3.81% of total), fixed = 1,724 (0.31 of total), total SNPs = 562,493), and West and South (significant = 17,117 (2.76% of total), fixed = 1,602 (0.26% of total), total SNPs = 621,184) (Table 3). It should be noted that the number of SNPs in each population comparison is influenced by differences in mapping coverage between the populations (Table 4, Table 3). However, divergence between *An. melas* South and the other populations was largest, whereas this population has the lowest number of mapped reads.

The *X* chromosome has a disproportionately large number of fixed and significant SNPs (Table 3, Figure 3) in both West - South and South - Bioko population comparisons. This pattern of elevated F_{ST} extends across the entire X chromosome (Figure 3). This could potentially be the result of increased genetic drift acting on polymorphisms due the lower effective population size of the *X* chromosome. Interestingly however, this *X* chromosome effect is not obvious between *An. melas* West and Bioko, the two most recently diverged groups.

We performed a gene ontology analysis on genes within windows that show evidence of non-neutral evolution (low Tajima's *D*). First we identified 100 kb sliding-windows with the lowest 5% Tajima's *D* values for each population (genome-wide, excluding heterochromatic regions) ($D < -0.200$, -0.096, and -0.148 for *An. melas* West, South, and Bioko, respectively). Next, we identified genes inside these windows that harbored SNPs with significant F_{ST} values in each pair-wise comparison. The West-South comparison yielded 95 significant SNPs located inside the exons of 64 genes. The molecular functions of these genes are associated with binding, catalytic activity, nucleic acid binding transcription factor activity, and receptor activity, among others (Table 6). The West-Bioko comparison yielded 79 significant SNPs located inside exons of 62 genes and the South-Bioko comparison yielded 188 significant SNPs located inside exons of 127 genes (Table 6). The molecular functions associated with these genes are

similar to those found in the West-South example. The most commonly found molecular functions (across all comparisons) include binding, catalytic activity, and nucleic acid binding transcription factor activity, and some genes are common among population comparisons (Table 6).

Common biological processes in all population comparisons include biological regulation, cellular processes, localization, and metabolic processes (Table 7). The South-Bioko comparison had 161 biological process gene ontology hits associated with the 127 genes in this analysis. The most frequent hits to protein classes across all comparisons were found in the hydrolase category, followed by proteases, nucleic acid binding proteins, proteases, and transcription factors (Table 8).

Our analysis of introgression between *An. melas* populations was based on the topology ((West, Bioko) South) *An. gambiae*) (Deitz *et al*., 2012), and screened for introgression between *An. melas* South and Bioko or South and West. This test found a genome-wide, positive deviation of the *D*-statistic (mean *D*-statistic = 0.040, Z-score = 21.80, Table 9), indicating an excess of ABBA sites and ancient or weak introgression betwen *An. melas* South and Bioko. An exception to this pattern was found on chromosome 2L (~22.25 Mb - 23.45 Mb), where *D*-statistic windows with a strong, negative deviation from zero (as low as -0.83) suggest recent *An. melas* South and West introgression (Figure 4). Interestingly, this introgression block overlaps precisely with a region of high nucleotide diversity in *An. melas* Bioko (Figure 3), and falls between the proximal breakpoint of the *2La* chromosomal inversion (which is fixed for the standard arrangement *in An. melas*) and the proximal breakpoint of the $2La^{2}$ chromosomal inversion (which is polymorphic within *An. melas*) (Coluzzii *et al.,* 2002, Sharakhov *et al.*, 2006, White *et al.,* 2007). The $2La^{2}$ inversion is specific to *An. melas* and is polymorphic within it (Coluzzi *et al.,* 2002). *An. melas* collected from Guinea Bissau and Cotonou, Benin (inside the range of the *An. melas* West cluster, Figure 1) share the standard arrangement ($2L+^{a2}$), while *An. melas* collected from Democratic Republic of the Congo (likely belonging to the *An. melas* South genetic cluster) are polymorphic for the standard and inverted arrangements ($2La^{2}$ & $2L+^{a2}$) (Coluzzi *et al.,* 2002).

Discussion

Population genomic analysis of *An. melas* West, South, and Bioko Island identified significant, genome-wide genetic differentiation, including the presence of numerous fixed SNPs throughout the genome in all *An. melas* population comparisons. Previous work based on microsatellites and mtDNA markers indicated levels of differentiation between *An. melas* forms that are on par, or exceed those observed between *An. gambiae* and *An. arabiensis* (Deitz *et al.,* 2012). Species pairs in the *An. gambiae* complex with comparable genetic differentiation are separated by strong pre- and post-mating isolation (Marchand 1983, Okereke 1980, Slotman *et al.,* 2004, Weetman *et al.,* 2014). Recently, the M and S molecular forms of *An. gambiae* were raised to species level (Coetzee *et al.,* 2013) based on well-documented ecological and some behavioral differences. These species have diverged considerably less than the three *An. melas* genetic clusters throughout most of their genomes but have several regions of high differentiation. This is not the case for the three *An. melas* forms, where, with the exception of a chromosome-wide *X* effect, genetic differentiation is distributed mostly evenly across the genome. This is consistent with a process of allopatric divergence with little gene flow/introgression. No evidence for "speciation islands", genomic regions with high levels of divergence that are maintained in the face of extensive hybridization gene flow (Turner *et al.,* 2005), was found in this study.

We used a simulation approach to construct an F_{ST} null distribution and false-discovery rate that incorporates both pool-size and sequencing coverage. To our knowledge, this is the first time this approach has been applied to a Pool-seq study. This allowed us to determine the F_{ST} significance threshold for each pair-wise population comparison. In doing so we assumed a starting allele frequency of 0.5, which results in the largest variance in the subsequent sampling steps of the simulation. In addition, we used a sequencing coverage of 30x for our simulations, which was the minimum sequencing coverage we required for F_{ST} calculations in our empirical analysis. Therefore, our approach is conservative. A downside of our approach is that it does not

provide q-values for individual SNPs, though our method could be adapted to do so in the future.

Intra-population nucleotide diversity in *An. melas* revealed remarkably similar patterns of variation across the genomes each population (Table 1, Figure 3). This shared pattern may be attributed to shared ancestry and genome organization (ex. chromosomal inversions). Additionally, selective constraints on many genes may be similar between these populations, as the ecology may be largely shared between forms. A single peak in nucleotide diversity on chromosome *2L* of *An. melas* Bioko is the exception. Interestingly, the results of the ABBA/BABA test suggested that this exact region introgressed between *An. melas* South and West (Figure 4). This highly surprising overlap suggests to us an alternative explanation: recent introgression of this region from *An. gambiae* (or more likely, the closely related *An. coluzzii,* see below), the outgroup species in the ABBA/BABA test, into *An. melas* Bioko. This would also create a pattern of BABA excess (suggesting introgression between *An. melas* South and West) and could explain the remarkably high nucleotide diversity in Bioko Island in this particular region. Both *An. coluzzi* and *An. melas* are present on Bioko Island (Overgaard *et al*., 2012), female hybrids between the two species are fertile (Davidson *et al.,* 1962), and extensive introgression between various species in the complex was recently documented (Fontaine *et al*., 2015). *An. gambiae s.s.* (i.e. *An. gambiae* S form) was eliminated from Bioko Island through a malaria control campaign, and only *An. coluzzii* (i.e. *An. gambiae* Forest-M form) remains (Overgaard *et al.,* 2012).

Genome-wide Patterson's *D*-statistic values from the ABBA/BABA test also suggests a slight bias toward a low level of ancestral introgression between *An. melas* South and Bioko (vs. between West and South). This finding is perhaps not surprising considering the geographical proximity of the *An. melas* South and Bioko populations used in this study (Ipono, Cameroon and Arena Blanca, Bioko Island, Equatorial Guinea, respectively) (Figure 1) in comparison to *An. melas* from Ballingho, The Gambia, which was our representative population of *An. melas* West.

Measures of nucleotide diversity in *An. melas* populations are less than half of the mean chromosomal nucleotide diversity values observed in *An. gambiae* (S form) populations collected from the north and south of Cameroon (0.008-0.15, Cheng *et al.,* 2012). This may reflect a lower *Ne* due to the patchy distribution of *An. melas* populations compared to *An. gambiae* (Athrey *et al.,* 2012, Deitz *et al.*, 2012). Genome-wide nucleotide diversity is the lowest in *An. melas* Bioko, which likely reflects a smaller effective population size (N_e) compared to the other *An. melas* populations. Previous findings also found that the Bioko Island population harbors lower levels of rarefied allelic richness at microsatellite loci, far fewer mitochondrial DNA haplotypes, and a much lower N_e compared to mainland populations (Deitz *et al.,* 2012). An alternative explanation of lower diversity due to founder effects is not supported by a previous Approximate Bayesian Computation analyses of the demographic history of these populations, which indicated that all three *An. melas* forms separated through vicariance events (Deitz *et al.,* 2012).

Mean chromosomal Tajima's *D* and nucleotide diversity were lowest on the *X* chromosome for *An. melas* South and Bioko (Table 1), and nucleotide diversity of the *An*. *melas X* chromosome was the second lowest of any chromosome arm. This may be due to positive selection on (partially) recessive alleles acting more strongly on the *X* chromosome. These findings are in agreement with an effects model (SnIPRE) analysis of natural selection between *An. melas* West, South, and Bioko Island populations, which found an increased selection effect of the *An. melas X* chromosome (Struchiner *et al.,* in review). Low diversity on the *X* chromosome of *An. melas* populations is consistent with findings in *An. gambiae s.s.* (Cohuet *et al.,* 2008, Holt *et al.*, 2002, Wilding *et al.*, 2009) and *An. arabiensis* (Marsden *et al.*, 2014). Introgression between member species of the *An. gambiae* complex is well documented (Fontaine *et al.*, 2015), but is limited between the *X* chromosome of *An. gambiae s.s.* and other members of the complex due to the *Xag* inversion, which covers ~60% of the *An. gambiae* s.s. *X* chromosome. The *Xag* inversion suppresses recombination between the *An. gambiae* and *An. arabiensis X* chromosomes, and plays a large role in their postzygotic reproductive

isolation (Slotman *et al.,* 2004, 2005b), preventing introgression. This suppressed introgression of the *X* chromosome between *An. gambiae* and *An. arabiensis* may have contributed to reduced nucleotide diversity on the *X* in these species (Marsden *et al*., 2014). Reduced introgression of the *X* may also contribute to its lower nucleotide diversity in *An. melas*, although its lower effective population size resulting in higher levels of genetic drift is probably a more important factor.

Mean Tajima's *D* was over three times lower in *An. melas* West as compared to the South and Bioko. As this is a genome-wide effect, it likely is the result of demographic factors, such as a recent population bottleneck in the *An. melas* West population analyzed. Windows of low Tajima's *D* are found throughout the genomes of the *An. melas* populations, which may indicate that these regions harbor genes under positive selection. Notably, very similar patterns of genome-wide Tajima's *D* are found in each *An. melas* population cluster. This suggests that while geographic isolation of *An. melas* clusters has greatly reduced gene flow between them, their resulting genetic differentiation is likely not a result of diverging selection pressures, which is expected to result in diverging Tajima's *D* patterns. The similar patterns of genome-wide Tajima's *D* likely also mean that genetic drift has not yet greatly impacted ancestral signatures of selection in these genomes.

Our gene ontology analysis explored the molecular and biological functions, and protein classes associated with genes found in low Tajima's *D* regions that also harbored significant or fixed SNPs. These included molecular functions associated with binding, catalytic, and nucleic acid binding transcription factor activity, biological functions including metabolic and cellular processes, localization and biological regulation, and protein classes such as enzyme modulators, nucleic acid binding, transcription factors, and transferases, among others (Tables 7-9). Future analyses of the functions of these genes might be able to reveal a link to their biological significance in *An. melas*.

Since early studies of host preference, parasitemia rate, and ecology of *An. melas* (Gelfand, 1955), and the original taxonomic, genetic, and descriptive studies of the *An. gambiae* complex (Davidson, 1962, White, 1974), *An. melas* has been considered a

malaria vector of minor importance due to its limited distribution and broad host preference. However, early studies focused on populations representing *An. melas* West alone. Recent studies have shown that on Bioko Island, Equatorial Guinea, *An. melas* populations readily feed on humans both indoors and outdoors (Reddy *et al.,* 2011), and are responsible for up to 130 malaria infectious bites/person/year in the village of Arena Blanca (Overgaard *et al.,* 2012). These studies highlight the important role that *An. melas* plays in malaria transmission. The results of this study, in combination with previous work (Deitz *et al.*, 2012), indicate that *An. melas* is undergoing an allopatric divergence process. Therefore, what we know about the ecology and behavior of *An. melas* West populations, which have been the focus of the handful of studies on the species (Bryan, 1983, Bryan *et al.,* 1987, Bogh *et al.,* 2007, Caputo *et al.,* 2008), may not hold true for the other *An. melas* forms. Additionally, as a member of a species complex that serves as a model for the speciation process, a better understanding of the population genomics of *An. melas* populations enhances our view of how the evolution of the *An. gambiae* species complex is influenced by the diverse host preferences, ecologies, distributions, and demographic histories of its member species.

CHAPTER II

DOSAGE COMPENSATION IN THE *ANOPHELES GAMBIAE* SPECIES COMPLEX

Introduction

Sex Determination in *Drosophila* and *Anopheles*

Insects have evolved a variety of mechanisms for sex determination. This is true even within mosquitoes (Diptera: Culicidae), where sex determination is controlled by a male-determining (M) factor located on the Y chromosome in species with heterogametic sex chromosomes (subfamily *Anophelinae*) (Krzywinska *et al.*, 2016) or on a homomorphic sex-determining chromosome (subfamily *Culicinae*) (Bachtrog *et al.,* 2014, Hall *et al.,* 2016). Until recently, M factors in mosquitoes had not been characterized because they are located in repeat-rich regions that were difficult to assemble into contigs. Transcription of the M locus during early embryo development in mosquitoes activates a cascade of sex-specific gene splicing that controls male and female development. The pathway by which this cascade takes place depends on the sex-determination system and can differ considerably between species (Gempe and Beye, 2011). However, the transcription factors *doublesex* (*dsx*) and *fruitless* (*fru*) are transcribed at the bottom of the sex determination pathway and are highly conserved among insects (Herpin and Schartl, 2015, Salz and Erickson, 2010, Wilkins 1995). Sex-biased splice isoforms of these genes have been found in mosquitoes, however it is unknown how their splicing is regulated (Biedler and Tu, 2016).

Drosophila melanogaster has a heterogametic (XX/XY) sex determination system that is dependant on X chromosome dosage (Cline, 1993). *Drosophila* males lack a M-locus, though their Y-linked genes are important for male-specific physiologies and behavior (Carvalho *et al.*, 2015, Carvalho *et al.*, 2000, Lemos *et al.*, 2008*,* Zhou *et al.,* 2012). In *Drosophila* females, a double dose of X chromosome-linked signal elements (XSE) initiates female-specific pre-mRNA splicing of *sex lethal* (*Sxl*) transcripts. SXL in turn regulates sex-specific splicing of *transformer* (*tra*), which, along with its non-sex-specific co-factor *transformer2* (*tra2*), promotes female-specific splicing of *dsx* and *fru*

pre-mRNAs. A single dose of XSE fails to initiate the cascade of female-specific splicing. As a result, male transcripts are produced and X chromosome dosage compensation takes place. DSX and FRU proteins are at the bottom of the sex-determination hierarchy, and male and female isoforms of DSX and FRU modulate the expression of genes involved in sexually dimorphic morphologies and physiologies. While male and female isoforms of DSX share the same DNA-binding domain, they differ in their C-terminal domains that confer sex-specific gene regulation (Clough *et al.*, 2014).

Mosquitoes in the subfamily Culicinae, including those in the genus *Aedes,* have homomorphic sex determination where sex-determining chromosomes appear similar to autosomes. Sex determination in *Aedes aegypti* is controlled by a dominant male-determining factor on the second chromosome that is "Y-like" in protein coding gene and repetitive element content (Newton *et al.* 1974). Using the chromosome quotient method of analyzing male vs. female gene expression during development, Hall *et al.* (2015) identified the gene *Nix* as the male determining gene in the M locus of *Aedes aegypti*. *Nix* is a potential splicing factor of *dsx* and *fru*. When Hall *et al* (2015) knocked out *Nix* in *Ae. aegypti* embryos, genetic males expressed a higher proportion of female *dsx* and *fru* isoforms, and exhibited feminization of external morphologies including antennae and copulatory structures.

Sex determination in anopheline mosquitoes (genus *Anopheles*) is controlled in a similar manner, despite the differences in sex-chromosome morphologies between *Anophelinae* and *Culicinae* (heteromorphic vs. homomorphic, respectively). The *Yob* gene is located within the Y-linked M-locus of *Anopheles gambiae sensu strictu. Yob* expression is male-specific, and starts within 2.5 hours of embryo oviposition (Krzywinska *et al.* 2016). Ectopic injections of *Yob* mRNA into embryos caused female lethality, but had no effect on genetic males. *Yob* activation occurs prior to *dsx* splicing by up to six hours, and is either direct or indirect upstream regulator of *dsx* splicing and sex determination in *Anopheles gambiae*.

Dosage Compensation in *Anopheles*

Dosage compensation in organisms with heterogametic sex determination overcomes the biological challenge "half dose" of X chromosome gene transcripts (or Z chromosome transcripts in ZZ/ZW systems) in the heterogametic sex. The need to equalize gene expression of a hemizygous X with that of autosomes is important because X chromosomes retain hundreds of functional genes that are actively transcribed in both sexes, while Y chromosomes are primarily heterochromatic and harbor relatively few genes. *D. melanogaster* mutants with triploid XXY sex chromosomes and diploid autosomes develop as infertile females due to a disruption of dosage compensation regulation. Dosage compensation and sex determination are not regulated independently; loss-of-function *sxl* mutations in *Drosophila* cause female embryonic lethality, likely due to the mis-regulation of dosage compensation (Biedler *et al*., 2016, Cline, 1978).

In contrast to *D. melanogaster*, an *Anopheles culicifacius* mosquito with triploid XXY sex chromosomes developed as an infertile male due to the presence of the M locus on the *Anopheles* Y chromosome (Baker and Sakai, 1970, Krzywinksa, 2016). Despite the differences in their mechanisms of sex determination (X chromosome dosage v. M locus, respectively) *Drosophila* and *Anopheles* both exhibit complete dosage compensation through the hyper-transcription of X chromosome genes in males to match (on average) expression levels of autosomal genes (Gelbart and Kuroda, 2009, Straub and Becker, 2007, Jiang *et al.,* 2015, Rose *et al*., 2016).

In *Drosophila*, dosage compensation via hyper-expression of the male X chromosome is tightly linked to sex determination by SXL (Lucchesi *et al.,* 2005, Gelbart and Kuroda, 2009, Bashaw and Baker, 1997). SXL is expressed only in females where it represses translation of the male-specific lethal 2 (msl2) mRNA, which controls dosage compensation in males. Mutations within *sxl,* or genes impacting *sxl* regulation result in over-expression of genes on the X chromosome and female lethality during embryogenesis (Erickson and Quintero, 2007, Lucchesi and Skripsky, 1981, Gergen, 1987, Hilfiker *et al.*, 1995). Ectopic expression of SXL in males disrupts dosage compensation and results in male death. *Yob* may influence the dosage compensation

pathway in *Anopheles gambiae,* which also relies on male hyper-expression of X-linked genes (Krzywinska *et al.*, 2016). Knockdown of *Yob* in embryos resulted in 100% male embryo mortality, potentially due to improper regulation of dosage compensation. Additionally, *Yob* injection into female embryos may induce dosage compensation (hyper-transcription of both copies of the X), resulting in their death due to abnormal X chromosome transcription (Krzywinska *et al.*, 2016).

One of two open reading frames of *Yob* analyzed by Krzywinska *et al.* (2016) showed evidence for purifying selection among members of the *Anopheles gambiae* species complex, indicating its conservation in the role of sex determination. However, no homolog of *Yob* was identified in *An. stephensi*, which diverged from *An. gambiae* ~30 million years ago (Hall *et al.* 2016, *Yob* is referred to as *YG2* in this article). *Yob* is the only protein-coding gene shared by the Y-chromosomes of all members of the *An. gambiae* species complex, supporting its role in sex determination (Hall *et al.* 2016). However, *Yob* shares high levels of plasticity with other protein coding genes and repeat sequences on the Y-chromosomes of members of the *An. gambiae* species complex. It is polymorphic in copy number between *An. gambiae* complex member species, and isolates of *An. gambiae s.s.*. Additionally, four haplotypes of the gene were present in one *An. gambiae s.s.* colony sequenced (Hall *et al.* 2016).

In *Drosophila*, high levels of Y-chromosome satellite DNA polymorphism between species have been implicated in hybrid incompatibility (Ferree and Barbash, 2009, Bayes and Malik, 2009, Satyaki *et al.,* 2014). *Hmr* and *Lhr* cause hybrid incompatibility between *D. melanogaster* and *D. simulans* and function within these species to suppress satellite DNA and transposable element transcription (Satyaki *et al.*, 2014). In *An. gambiae*, satellite sequences AgY477 and AgY373 are expressed exclusively in male testes, but are absent or structurally modified in other species in the complex (Hall *et al.* 2016).

Previous work concluded that *Anopheles stephensi* mosquitoes exhibit complete dosage compensation (Jiang *et al.*, 2015), and the same is true for *Anopheles gambiae* (Rose *et al.,* 2016). Like *Drosophila,* dosage compensation is accomplished through the

hyper-transcription of non sex-specific or -biased genes on the male homomorphic X chromosome to reach similar expression levels of autosomal loci, and to roughly equalize expression between the sexes (Rose *et al.*, 2016). However, the high level of protein coding gene, satellite DNA, and *Yob* haplotype turnover between members of the closely related *An. gambiae* complex suggests that further insight into these differences in Y-chromosome content may yield insight into the reproductive isolation of these species. To this end, I have addressed the following questions in this chapter:

1] Does dosage compensation occur in *Anopheles gambiae* complex member species *An. arabiensis* and *An. quadriannulatus* as it does in *An. gambiae s.s.*?

2] Does hybridization between member species of the *An. gambiae* complex result in a disruption of dosage compensation?

3] Does the direction of hybridization influence dosage compensation?

To address these questions, I analyzed male vs. female and X chromosome vs. autosome gene expression levels in *An. coluzzii* (formerly *An. gambiae s.s.* M form), *An. arabiensis*, and *An. quadriannulatus*, and bi-directional hybrids between *An. coluzzii* and *An. arabiensis,* as well as *An. coluzzii* and *An. quadriannulatus.* By analyzing both cross directions of each species comparison, I was able to address whether or not trans-interactions between the hemizygous X chromosome and the heterospecific set of autosomes results in the disruption of dosage compensation in males.

Methods

Mosquito Rearing

The mosquitoes used in this experiment were derived from the SUA2La (*An. coluzzii*), SANGQUA (aka SANGWE, *An. quadriannulatus*), and DONGOLA (*An. arabiensis*) strains, and bi-directional crosses between *An. coluzzii* (COLZ) and *An. arabiensis* (ARAB)*,* and *An. coluzzii* and *An. quadriannulatus* (QUAD). Cross directions are indicated herein female x male (COLZxARAB, ARABxCOLZ, COLZxQUAD, QUADxCOLZ). Adult mosquitoes were reared in 12 hr light / 12 hr dark cycles at 25 ˚C and 70-80% relative humidity. Adults had constant access to 5% sucrose solution. The

colonies were blood fed via an artificial membrane weekly with warmed, defibrinated sheep blood (Hemostat Laboratories, Dixon, CA). Adults were provided wetted filter paper during the period of 48-72 hr post-feeding to allow females to oviposit. Eggs were placed in bins of de-ionized water 0-12 hr post-oviposition, and emerging larvae were fed Tetra brand TetraColor Tropical Crisps (fish food) *ad libitum* until pupation. Mosquitoes were collected 12 hr post-pupation with an eyedropper, sexed twice, and placed five each into 1.5 mL micro-centrifuge tubes. At this point the pupae were alive and swimming in de-ionized water.

Pure strain *An. coluzzii, An. arabiensis*, and *An. quadriannulatus* pupae were isolated from lab colonies as outlined above. Bi-directional crosses were performed by isolating 100-200 pupae from the respective colonies, sexing the pupae twice, and allowing the male and female pupae from both parental strains of the cross to emerge as adults in a common adult cage. This approach ensures that both male and female mosquitoes are virgins. Each cross was comprised of 50-100 adults of each sex/species. Sexing of pupae was performed twice to ensure that the proper sex of each species was introduced into the cross. Rearing, feeding, and egg laying conditions were the same as for pure strain/species mosquitoes described above, as was the rearing and isolation of F1 hybrid pupae for RNA extraction.

RNA Isolation and Sequencing

After isolating pupae, the 1.5 mL micro-centrifuge tubes were placed on ice and taken to an RNA-only extraction bench. Here, the water was drawn from the micro-centrifuge tubes using a Pasteur pipette, and the pupae were immediately ground with a disposable pestle in Qiagen Buffer RLT (lysis buffer). Total RNA extractions were performed immediately using a Qiagen RNeasy Mini kit (Qiagen, Redwood City, CA) with an on-column DNase treatment according to the manufacturers protocol. After an extraction was completed a 5.0 uL aliquot was removed for RNA quantification and quality assessment. The remaining extracted RNA was stored at -80˚C.

Each extraction product was run on an Agilent Bioanalyzer to assess quality, and on a NanoDrop spectrophotometer to quantify the RNA. RNA extraction products that

passed quality control were used to form two biological replicates for both sexes of each mosquito pure strain and cross. Biological replicates are comprised of pooled, equimolar amounts of RNA from two or four extraction products (10 or 20 total mosquito pupae). Pooled RNA samples were again run on the Agilent Bioanalyzer and NanoDrop spectrophotometer prior to library prep and sequencing to ensure that the samples had not degraded during the pooling process. Library prep was performed using an Illumina TruSeq RNA library prep kit (Illumina Inc., San Diego, CA). In total, 28 libraries were sequenced on seven lanes of Illumina HiSeq 2500 machine using 125 bp single-end chemistry.

Pseudo-Genome Construction, RNAseq Alignment, and Transcript Abundance

I used the FastQC package version 0.11.4 (Babraham Bioinformatics) to visualize RNAseq run and read quality. Illumina TruSeq adapters were identified and removed using Trimmomatic version 0.30 (Bolger *et al.,* 2014), while simultaneously soft-clipping the reads from both 5' and 3' ends to an average phred quality score of 20, with no single bp in a four bp window below a phred quality score of 10. Only reads ≥ 50 bp were retained for alignment and subsequent analyses.

Trimmed RNAseq data from parental species were initially aligned to their respective, species-specific genomes, which were obtained from VectorBase.org (AcolM1 for *An. coluzzii,* AaraD1 for *An. arabiensis*, and AquaS1 for *An. quadriannulatus*, Giraldo-Calderon *et al*., 2015). Initial RNAseq alignments were performed using the splice-aware alignment software STAR version 2.4.2a (Dobin *et al*., 2012). Initially, I followed the Genome Analysis Toolkit (GATK) best practices guide for calling variants from RNAseq data (https://software.broadinstitute.org/gatk/guide/article?id=3891), which employs a STAR 2-pass protocol. In short, parental reads were first aligned to their respective species-specific reference genomes using STAR. Next, an index is created for each alignment, which annotates all splice junctions found during the first alignment round. This index informs the second alignment with STAR, which re-aligns the reads around splice junctions. Next, I used the Picard tools version 2.8.1

(http://broadinstitute.github.io/picard/) "MarkDuplicates" function to mark and remove reads derived from PCR duplication during library prep.

BAM alignment files from all individuals of each parental species (biological replicate one and two of both males and females) were then merged using Picard MergeBamAlignment, and the SAMtools/BCFtools version 1.3 (Li *et al.*, 2009) mpileup/call function was used to call variants within each merged alignment. This process was informed by a prior mutation rate of 0.02. Next, I used the GATK VariantFiltration tool to select SNPs only from the variant call files of each parental species. The vcfutils.pl varFilter function, a component of the SAMtools package, was used to filter SNPs using default parameters with the exception of not filtering for strand bias, as this can be expected in RNAseq data. Next, filtered SNPs for each species were converted to the genomic coordinates of the *An. gambiae s.s.* PEST strain AgamP4 genome (VectorBase.org). The coordinate lift-over process was performed using the Picard LiftoverVCF function, using a chain file to match base-pair coordinates between genomes. Chain files were created from pairwise multiple-alignment files between the *An. coluzzii, An. arabiensis*, and *An. quadriannulatus* genomes and the *An. gambiae s.s.* PEST reference. These multiple alignment files were originally published by Fontaine *et al.* (2015) and were downloaded from VectorBase.org. Once species-specific SNPs were converted to AgamP4 coordinates, they were incorporated into the AgamP4 genome using the GATK FastaAlternateReferenceMaker function to create a "pseudo-genome" for each species using the AgamP4 backbone/coordinates. By incorporating SNPs from each species into the AgamP4 genome, I was able to use the AgamP4 genomic coordinates and gene set as a common framework to quantify transcript abundance within each sample.

Next, I again performed a STAR 2-pass alignment for each parental library using the species-specific AgamP4 pseudo-genome as a reference. SNPs were called using the methods described above, and were then incorporated into the previous psuedo-genome reference to further populate the AgamP4-based pseudo genome of each species with variation from that strain. The second-round, AgamP4-based pseudo-genomes of each

species were then used as a reference to create a cDNA pseudo-genome for each species. cDNA genomes were created using the "generate_transcripts" function of rSeq version 0.2.2 (Jiang and Wong, 2009). This function uses the start and end positions of each gene exon in a given gene annotation file to pull cDNA sequences from the input genome. Each transcript is then written to a FASTA genome file. Gene transcript coordinates were derived from the AgamP4.4 gene set (VectroBase.org).

Parental species libraries (*An. coluzzii, An. arabiensis,* and *An. quadriannulatus*) were aligned to their respective cDNA pseudo-genomes using the "--very-sensitive" mode of Bowtie 2 version 2.2.9 (Langmead and Salzberg, 2012). Only uniquely mapped reads were retained to calculate transcript abundances. Transcripts abundances (read counts per gene) were quantified using the SAMtools "idxstats" function, which reports the number of reads in a SAM/BAM alignment file that are aligned to each contig of the reference sequence. In the case of the parental cDNA pseudo-genomes, each contig is a gene transcript. So, this function is counting the number of transcripts aligned to a gene.

RNAseq reads derived from the bi-directional crosses (males and females of COLZxARAB, ARABxCOLZ, COLZxQUAD, QUADxCCOLZ) (hereafter F1 hybrids) were aligned to diploid pseudo-genomes comprised of the AgamP4.4 based cDNA pseudo-genomes of each parental species. In short, the "generate_transcripts" function of rSeq version 0.2.2 (Jiang *et al.*, 2009, Salzman *et al.,* 2011) was used to create cDNA genomes from the AgamP4 based pseudo-genomes of each parental species (as above). The maternal and paternal cDNA genomes for each hybrid were then merged into a diploid genome using the merge_pat_mat_fasta.pl script of the ASE-TIGAR program (Nariai *et al.*, 2016, http://nagasakilab.csml.org/ase-tigar/), which essentially concatenates the cDNA genomes of the maternal and paternal strains into one genome file. RNAseq reads derived from F1 hybrids were assessed for quality and trimmed using FastQC and Trimmomatic as described above. Reads were mapped to the diploid, bi-parental cDNA pseudo-genomes for the respective cross using the "--very-sensitive" mode of Bowtie 2 version 2.2.9 (Langmead and Salzberg, 2012). Each F1 hybrid read was allowed to align up to 100 times across the cross-specific diploid reference genome.

F1 hybrid BAM alignment files were then processed using the program ASE-TIGAR to identify the parent of origin of each transcript. ASE-TIGAR uses a Bayesian approach to estimate transcript abundance from RNAseq data that has been aligned to diploid pseudo-genomes. It does this by modeling the haploid choice as a hidden variable, and simultaneously estimates isoform abundance by variational Bayesian inference. ASE-TIGAR uses variation in each pseudo-genome to appropriately match each transcript to its parent of origin when fixed genetic differences exist between the parental species. Transcripts for which the parent of origin cannot be determined due to a lack of genetic differentiation between them are assigned to a parent randomly. With this approach, we are able to quantify total expression (total transcript abundance) and allele-specific expression in F1 hybrids. Total transcript abundance (maternal + paternal expression in F1 hybrids) was used for the subsequent analysis of dosage compensation.

Analysis of Dosage Compensation

Transcript abundances for each gene within a sample were standardized by converting them to RPKMs (reads per kilo-base of transcript per million reads) in R (www.cran.r-project.org, R Core Team, 2013). This allows the comparison of gene expression across samples. RPKM was calculated by dividing the number of reads per gene by the kilobase length of the gene's total exonic region, and then by the total number (in millions) of reads in the sample.

I employed two methods to assess dosage compensation in parental and F1 hybrid samples. First, I compared the ratio of median expression of X-linked and autosomal genes in each sample. This was performed by correcting the RPKM value of each X-linked gene by the median RPKM expression level of all autosomal loci. I used the "boot" package in R (Davison and Hinkley, 1997, Cantly and Ripley, 2015) to calculate 95% confidence intervals of the median of this distribution, for each sample, using 10,000 replicates. I analyzed X to autosomal (X:A) median gene expression as a function of increasing minimum RPKM values between zero and 10. In order to be included in the analysis at a specific RPKM threshold, a gene had to exceed the specified RPKM level in all samples in a species comparison (ex. *An. coluzzii*, *An. arabiensis*, and

their hybrids). Because of this, the number of genes included in this analysis differs for *An. coluzzi* between the *An. coluzzi* - *An. arabiensis* and the *An. colluzzi* - *An. quadriannulatus* species comparisons. An analysis of X:A median gene expression with no RPKM cut-off (RPKM ≥ 0.0) includes all transcriptionally active and inactive genes. The incorporation of non-active genes into analyses of dosage compensation can bias X:A median gene expression ratios if inactive genes are non-randomly distributed between the X chromosome and autosomes (Kharchenko *et al.*, 2011). Thus, by analyzing X:A expression ratios at increasing minimum RPKM thresholds (RPKM > 0.0, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0), I was able to identify the minimum RPKM level that removes bias during subsequent gene expression analyses.

By comparing X:A median gene expression ratios between males and females, we can better understand if genes on the male hemizygous X chromosome are hyper-expressed to match the level of X chromosome expression in females. In females, we expect X:A expression ratios to be near one. We do not expect the X:A expression ratio to differ between males and females if the X chromosome is hyper-expressed in males. If dosage compensation is not active and the X is not hyper-expressed in males, we would expect a X:A expression ratio of males to behalf of female expression. X:A expression ratios in males and females were compared to the expression ratio of the second and third chromosomes (2:3). We expect the 2:3 expression ratio to be equal in both sexes. Thus, 2:3 median expression serves as an internal standard with which to compare X:A median expression. 2:3 expression ratios were calculated by correcting the RPKM value of each chromosome 2 gene by the median RPKM level of all chromosome 3 genes. 2:3 expression ratios were also analyzed at increasing minimum RPKM thresholds (RPKM > 0.0, 0.2, 0.5, 1.0, 2.0, 5.0, 10.0).

I used the "boot" package in R (Davison and Hinkley, 1997, Cantly and Ripley, 2015) to calculate the 95% confidence intervals of the median of the expression ratio distribution, for each sample, using 10,000 replicates. A Kruskal-Wallis test was used to test for significant differences in the median of the X:A expression ratio distribution *between* each sample in a species comparison (*An. coluzzii* - *An. arabiensis* or *An.*

coluzzii - *An. quadriannulatus*) (Kruskal and Wallis, 1952). Additionally, an ANOVA was used to test for significant differences in the mean of the X:A expression ratio distribution *between* each sample in a species comparison. The same tests were performed to analyze significant differences between the distributions of 2:3 expression ratios between samples. Statistical tests were performed to compare the expression ratio distributions between samples at each minimum RPKM cut-off.

The Kruskal-Wallis test (Kruskal and Wallis, 1952) and an ANOVA were used to test for significant differences between the X:A and 2:3 expression ratio distributions *within* a sample, at each minimum RPKM cut-off. Where appropriate, a Tukey's post-hoc test (Tukey, 1949) was performed to test for significant differences in the means of pair-wise comparisons, and a Dunn's test (Dunn, 2012) was used to test for significant differences in the medians of pair-wise comparisons.

Next, I analyzed the male to female (M:F) expression ratios of X-linked and autosomal genes. Genes were only included in this analysis if they exceeded a RPKM of 10.0 across all samples in a species comparison (ex. *An. coluzzii*, *An. arabiensis*, and their hybrids). The null expectation of this approach is that the M:F ratio of X-linked and the M:F ratio of autosomal genes will not differ if dosage compensation is operating effectively. M:F expression ratios were calculated by first, calculating the mean RPKM expression of each gene between the two biological replicates of each sex of each parental strain / F1 hybrid (ex. (QUAD male 1 + QUAD male 2) / 2). Then, the male to female expression ratio was calculated for each gene within a parental strain or F1 hybrid. The 95% confidence intervals of the median of each M:F expression ratio distribution were calculated with "boot" package in R (Davison and Hinkley, 1997, Cantly and Ripley, 2015) using 10,000 bootstrap replicates. Medians and 95% confidence intervals were calculated for X-linked and autosomal distributions separately. I used an ANOVA to test for significant differences between the means of the X-linked, M:F expression ratio distribution and the autosomal M:F expression ratio distribution within each parental species / F1 hybrid. Similarly, a Kruskal-Wallis test was used to test for significant differences between the medians of the X-linked, M:F expression ratio

distribution and the autosomal M:F expression ratio distribution within each parental species / F1 hybrid.

Results

RNA Sequencing and Alignment

The sequencing effort yielded over 1.88 billion reads that were uniquely assigned to a library based upon their Illumina TruSeq barcode. The average number of reads per sample is 67.23 million, though this number is highly skewed by one library COLZ x QUAD Female 2) that had 248.18 million reads (Tables 10 and 11). If COLZ x QUAD Female 2 is removed, the mean number of reads per sample is 60.53 million, with a range of 51.73 to 75.01 million. The mean mapping efficiency of parental libraries to their respective, species-specific reference genomes was 85.5%, with a minimum of 80.9% (COLZ male 2) and a maximum of 88.7% (ARAB male 1, Table 10). These libraries were combined for all samples within a parental species, SNPs were called for each and were converted to the coordinates of the *An. gambiae* PEST strain genome (AgamP4), and were then incorporated into the AgamP4 genome to create a pseudo-genome for each parental species. Mapping efficiency of all parental libraries increased when aligned to their respective AgamP4-based pseudo-genomes (mean 97.3%, min. 95.2%, max 99.1%, Table 1). This is likely because the AgamP4 genome is more complete, and is assembled into chromosomes, whereas the *An. coluzzi, An. arabiensis*, and *An. quadriannulatus* genomes are comprised of thousands of scaffolds.

After calling SNPs from the combined pseudo-genome based alignments of each parental species, these were again incorporated into the AgamP4-based pseudo-genome for each species. cDNA pseudo-genomes for each species were then constructed, and parental libraries were aligned. The mapping efficiency of all libraries to their respective cDNA pseudo-genomes was reduced in comparison to either DNA genome alignment (mean 63.8%, min. 62.3%, max 65.2%, Table 1). However, this reduction was uniform across each species. This reduction is may be because the AgamP4 gene set that the cDNA genomes were based off of did not include gene models for all transcribed genes

in the RNAseq datasets. F1 hybrid RNAseq reads were aligned to the respective bi-parental cDNA pseudo-genome for that cross. After filtering BAM alignments with ASE-Tigar, the average mapping efficiency for F1 hybrid libraries was 64.3% (min. 62.3%, max. 66.7%, Table 11). Thus, the mapping efficiency of F1 hybrid and parental libraries to the cDNA pseudo-genomes was equivalent.

Chromosome Ratios - *An. coluzzi* - *An. arabiensis* Comparison

In this section I report X to autosomal and chromosome 2: 3 gene expression ratios between *An. coluzzi* and *An. arabiensis* males and females, and their F1 hybrids. It is important to understand if dosage compensation acts within each species, and if it differs between the parental species, because this informs our interpretation of the F1 hybrid results. If X to autosomal and chromosome 2:3 gene expression is comparable between parental species and their hybrids, we can conclude that dosage compensation is not disrupted during hybridization.

Including genes with low expression levels can bias the analysis of dosage compensation, so I analyzed X to autosomal and chromosome 2: 3 gene expression ratios at increasing minimum RPKM thresholds from zero to 10. The number of genes included in each dataset is reported in Table 12. The means of the X:A and 2:3 expression ratios distributions can be skewed by a few genes with very low expression relative to the median gene expression of the denominator chromosome(s) (ex. very low expression of a X-linked gene relative to the median expression of autosomal loci). As minimum RPKM thresholds increase, these genes are removed from the analysis, and the mean better represents the center of the distribution (mean and medians converge). For this reason, I tested for significant differences between the means and medians of the X:A and 2:3 distributions at each RPKM threshold. It is important to note that as the minimum RPKM threshold increases, the number of genes included in the analysis decreases (Table 12).

As genes with low expression levels are removed from the dataset, the mean and median X:A expression ratios are near 1.0 for males and females of pure species and F1 hybrids. The median values of the X:A expression ratio distributions and the 95%

confidence intervals of these values are reported for each sample and minimum RPKM threshold for the *An. coluzzii* - *An. arabiensis* comparison (parents and F1 hybrids) in Table 13. The means of the X:A expression ratio distributions among these samples differ significantly (ANOVA p-value > 0.05) up to a minimum RPKM > 1.0, while the medians of these distributions only differ significantly at a minimum RPKM > 0.0 (Kruskal-Wallis p-value = 0.036, Table 13). As expected, the medians of these distributions converge prior to the means due to the skewed distribution. Median X:A expression ratios range from 0.71 to 1.13 across all samples and minimum RPKM cut-offs. However, as the RPKM cut-off increases, all median X:A expression ratios increase and are closest to 1.0 at the minimum RPKM > 5.0 or 10.0. At a minimum RPKM > 10.0, the 95% confidence intervals of the median for all samples includes 1.0 (Table 13, Figure 5).

The means of the 2:3 expression ratio distributions among these samples differ significantly (ANOVA p-value > 0.05) at all minimum RPKM values, suggesting that a small number of chromosome 2 genes with very high expression levels relative to the median expression of chromosome 3 genes are skewing the distributions (Table 14). However, the medians of the 2:3 expression ratio distributions do not differ significantly at a minimum RPKM cut-off of 10.0 (Kruskal-Wallis p-value = 0.749, Table 14). Median 2:3 expression ratios range from 0.94 to 1.18 across all samples and minimum RPKM cut-offs. As the minimum RPKM increases, median 2:3 expression ratios generally decrease and are closest to 1.0 at the minimum RPKM > 5.0 or 10.0. At a minimum RPKM > 10.0, the 95% confidence intervals of the median for all samples include 1.0 (Table 14, Figure 5). Additionally, at a minimum RPKM > 10.0, the 95% confidence intervals of X:A and 2:3 ratios within every sample overlap, and both 95% confidence intervals include 1.0 (Tables 13,14, Figure 5).

A pairwise analysis of male and female samples within a species or cross in the *An. coluzzii* - *An. arabiensis* comparison (for example, pair-wise comparisons between male and female biological replicates of COLZ x ARAB F1 hybrids) found that only the mean 2:3 expression ratios of two comparisons (ARABxCOLZ male 1 & ARABxCOLZ

female 1, and ARABxCOLZ male 1 & ARABxCOLZ female 2) differed significantly at a minimum RPKM > 10.0 (Table 15). None of the median X:A or 2:3 expression ratios differed significantly from each other in any pairwise comparisons above a minimum RPKM > 5.0 (Table 16). This suggests that, in agreement with previous findings from *An. gambiae* (Rose *et al.,* 2016), X:A and 2:3 gene expression ratios are equivalent between males and females when only transcriptionally active genes are included in the analysis.

A test for significant differences between the means (Table 17) and medians (Table 18) of the X:A and 2:3 expression ratio distributions within each sample found no significant differences between the means within a sample at or above a minimum RPKM > 2.0, and no significant differences between the medians within a sample above at a minimum RPKM > 10.0. At a minimum RPKM > 5.0, the medians of the X:A and 2:3 expression ratio distributions differed significantly only in one sample (Table 18). Thus, at a minimum RPKM > 10.0, the medians of the X:A gene expression distributions did not differ significantly between any samples (Tables 13,16), the medians of the 2:3 gene expression distributions did not differ significantly between any samples (Tables 14, 16), the 95% confidence intervals of the medians for both ratios included 1.0 in all samples (Tables 13,14, Figure 5), and the medians of the X:A and 2:3 expression ratio distributions did not differ significantly within samples (Table 18). This data shows that in *An. coluzzi, An. arabiensis*, and their F1 hybrids, dosage compensation is acting to balance X and autosomal gene expression in hemizygous males, and dosage compensation is equivalent between the parental species.

Chromosome Ratios - *An. coluzzi - An. quadriannulatus* Comparison

In this section I report X to autosomal and chromosome 2: 3 gene expression ratios between *An. coluzzi* and *An. quadriannulatus* males and females and their F1 hybrids. Again, X to autosomal and chromosome 2 to 3 gene expression ratios were analyzed at increasing minimum RPKM thresholds from zero to 10. The number of genes included in each dataset is reported in Table 19 and don't differ drastically from the *An. coluzzi - An. arabiensis* comparison. As above, I tested for significant differences

between the means and medians of the X:A and 2:3 distributions at each RPKM threshold.

The median values of the X:A expression ratio distributions and the 95% confidence intervals of these values are reported for each sample and minimum RPKM threshold for the *An. coluzzii* - *An. quadriannulatus* comparison in Table 20. The means of the X:A expression ratio distributions among these samples do not differ (ANOVA p-value > 0.05) at any minimum RPKM value. The medians of these distributions differ significantly at all minimum RPKM levels with the exception of minimum RPKM > 10.0 (Kruskal-Wallis p-value = 0.16, Table 20). Median X:A expression ratios range from 0.70 to 1.09 across all samples and minimum RPKM values. As in the *An. coluzzii* - *An. arabiensis* comparison, as the minimum RPKM increases, median X:A expression ratios generally increase and are closest to 1.0 at the minimum RPKM > 5.0 or 10.0 (Table 20, Figure 6). At a minimum RPKM cut-off of 10.0, the 95% confidence intervals of the median for all samples include 1.0 (Table 20, Figure 6). This demonstrates that the male hemizygous X is hyper-expressed to match autosomal expression in both species and both directions of the hybrid cross.

The means of the 2:3 expression ratio distributions samples differ significantly (ANOVA p-value > 0.05) at all minimum RPKM values in the *An. coluzzi* - *An. quadriannulatus* comparison (Table 21). The medians of the 2:3 expression ratio distributions do not differ significantly at a minimum RPKM > 0.5, 1.0, or 5.0, but are significantly different at RPKM > 0.0, 0.2, 2.0, and 10.0 (Table 21). Median 2:3 expression ratios range from 0.93 to 1.17 across all samples and minimum RPKM cut-offs. As the RPKM cut-off increases, median 2:3 expression ratios generally decrease and are closest to 1.0 at the minimum RPKM > 2.0, 5.0, or 10.0 (Figure 6). At minimum RPKM > 5.0 and 10.0, the 95% confidence intervals of the median for all samples include 1.0 (Table 21, Figure 6). At a minimum RPKM > 10.0, the 95% confidence intervals of the median of X:A and 2:3 expression ratios overlap within all samples, and both 95% confidence intervals include 1.0 (Tables 20, 21, Figure 6).

A pairwise analysis of male and female samples within a species or cross in the *An. coluzzii* - *An. quadriannulatus* comparison (for example, pair-wise comparisons between male and female biological replicates of COLZ x QUAD F1 hybrids) found that none of the mean 2:3 expression ratios differed significantly at a minimum RPKM > 5.0 or 10.0 (Table 22). Additionally, none of the median X:A or 2:3 expression ratios differed significantly from each other in any pairwise comparisons above a minimum RPKM of 0.5 (Table 23).

A test for significant differences between the means of the X:A and 2:3 expression ratio distributions within each sample found no significant differences for each sample at or above a minimum RPKM > 2.0 (Table 24). At a minimum RPKM > 10.0, one sample (COLZxQUAD Female 1) had a statistically significant difference between the medians of its X:A and 2:3 expression ratio distributions (Kruskal-Wallis p-value = 0.036, Table 25), suggesting that these distributions differ.

Thus, at a minimum RPKM > 10.0, the medians of the X:A expression ratio distributions did not differ significantly between any samples (Tables 20, 22). While the medians of the 2:3 gene expression distributions do differ significantly between all samples in *An. coluzzii* - *An. quadriannulatus* species comparison at a minimum RPKM > 10.0 (Table 21), pairwise analyses among biological replicates and sexes within each species / hybrid (Table 23) shows that this result is driven by comparisons of medians that are not relevant to the question of dosage compensation (i.e. comparing an *An. coluzzii* male to an *An. quadriannulatus* female (Table 23)). This suggests that dosage compensation is occurring in each species/hybrid, and equivalent between each.

No significant differences in the medians of X:A or 2:3 expression ratio distributions were found in pairwise comparisons between biological replicates and sexes within a species / hybrid at a minimum RPKM > 0.2 or higher (Table 23). This indicates that the male hemizygous X is being expressed at same level as the female X in *An. coluzzi, An. quadriannulatus,* and their hybrids. The 95% confidence intervals of the medians for both X:A and 2:3 expression ratios included 1.0 in all samples at a minimum RPKM > 10.0 (Tables 20, 21, Figure 6), and the medians of the X:A and 2:3

expression ratio distributions did not differ significantly within most samples at a minimum RPKM > 10.0 (Table 25). The one exception was a female, where we know dosage compensation does not operate in the first place. In this sample (COLZ x QUAD Female 1) the median X:A expression ratio was 1.03 (Table 20), and the median 2:3 expression ratio was 1.00 at a minimum RPKM > 10.0 (Table 21). While these values differ significantly, they are both very close to 1.0.

Male to Female Expression Ratios

Male to female expression ratios of X-linked and autosomal genes were compared between parental species and hybrids in the *An. coluzzi* - *An. arabiensis* and *An. coluzzi* - *An. quadriannulatus* comparisons (parental and F1 hybrids). Based on the findings of the chromosome ratio analyses (above), only genes with a minimum RPKM > 10.0 across all samples in a species comparison were included in this analysis. Median M:F expression ratios range from 0.87 to 1.07 for X-linked loci, and from 0.90 to 1.11 for autosomal loci (Table 26, Figures 9, 10). The **means** of the X-linked and autosomal M:F expression ratio distributions did not differ significantly **within** any species or hybrid. However, the **medians** of the X-linked and autosomal M:F expression ratio distributions differed significantly **within** all species / hybrids (Table 26). In all cases, the median of the M:F expression ratio distribution was lower for X-linked genes vs. autosomal genes (Figures 9, 10), indicating a higher proportion of female-biased genes expressed on the X chromosome compared to the autosomes. An additional explanation for this pattern (though not mutually exclusive) is a higher proportion of male-biased genes on the autosomes compared to the X chromosome, which is to be expected.

Discussion

This study represents the first analysis of dosage compensation in the closely related species *An. coluzzii*, *An. arabiensis*, and *An. quadriannulatus.* The members of the *An. gambiae* complex were thought to be a single species until experimental crosses between five 'strains' revealed the presence of hybrid sterility in F1 males **(**Davidson 1962, White 1974, Hunt *et al.* 1998). Later studies described additional species in the

complex to bring the current total to eight species (Coetzee *et al.*, 2013). Previous work focused on post-zygotic isolation in the complex identified loci in *An. gambiae* that are responsible for male and female sterility when introgressed into an *An. arabiensis* genetic background, and vice versa, by conducting QTL analyses of sterility in (*An. gambiae* x *An. arabiensis*) x *An. arabiensis* and (*An. gambiae* x *An. arabiensis*) x *An. gambiae* backcrosses (Slotman *et al.* 2004, 2005). Genes located on the X chromosome had a large effect on the hybrid sterility phenotype in both directions of the cross. Additionally, Slotman *et al.* (2005) identified interacting factors between the *An. gambiae* X chromosome and both *An. arabiensis* autosomes that cause complete inviability. This information, coupled with more recent knowledge of Y-chromosome divergence between member species of the *An. gambiae* complex in protein coding gene content, haplotype diversity and copy number of the *Yob* sex-determination locus, and repeat sequence copy number and diversity (Hall *et al*, 2016), led to the question of whether or not dosage compensation is disrupted in F1 hybrid males.

Previous studies in *Drosophila* analyzed dosage compensation across closely related species and found that hybridization does not have adverse effects on dosage compensation (Barbash, 2010). However, due to the different mechanisms of sex-determination between *Drosophila* and *Anopheles* (dosage vs. the *Yob* M-locus), the novel evolution of dosage compensation within the *Culicidae* (Biedler and Tu, 2016), and the rapid evolution of the M-locus harboring Y chromosome in the *Anopheles gambiae* species complex (Hall *et al*., 2016), I explored whether these differences impacted the hybrid dosage phenotype in the *Anopheles gambiae* complex.

My analysis of X chromosome to autosome, and second to third chromosome gene expression ratios provides further evidence for complete dosage compensation in the *An. gambiae* species complex. Specifically, this analysis has shown that genes on the male hemizygous X chromosome hyper-expressed to match the expression of autosomal genes, and on average, the expression level of X chromosome genes in females.

In pure species *An. coluzzi, An. arabiensis,* and *An. quadriannulatus,* median X:A expression ratios are near 1.0 in both sexes of each species at all minimum RPKM

thresholds. The same is true for the F1 hybrids. This indicates that dosage compensation is occurring in each species, and that the mechanisms controlling dosage compensation are not disrupted through hybridization. At a minimum RPKM > 10, median X:A and 2:3 expression ratios do not differ significantly between the sexes of any pure species or their hybrids (Tables 16, 23). With one (hybrid female) exception, median X:A and 2:3 expression ratios do not differ significantly within male or female samples, irrespective of their species or cross (Tables 18, 25). While median male to female ratios of X-linked and autosomal genes in this study differ significantly, both ratios are near one in all species / hybrids, and are in agreement with prior work on dosage compensation *An. gambiae s.s.* and *An. stephensi*.

Previous studies focused on *An. stephensi* and *An. gambiae s.s* had shown evidence for complete dosage compensation using similar methods of comparing chromosome ratios and analyzing male to female expression ratios (Jiang *et al*., 2015, Rose *et al.*, 2016). Jiang *et al*. (2015) reported median X:A expression ratios in *An. stephensi* males and female between 0.92 and 0.98 at minimum RPKM thresholds between zero and four, with no significant differences between median X and autosomal expression in males or females above a minimum RPKM > 2.0. They also compared expression ratios between X-linked genes and their one-to-one orthologs in *Ae. aegypti,* which does not have heteromorphic sex chromosomes. The median RPKM ratio of X-linked *An. stephensi* genes to their *Ae. aegypti* orthologs was close to one in both sexes after normalization.

Rose *et al.* (2015) analyzed dosage compensation in *An. gambiae s.s.* (M form) larvae and pupae, finding evidence for complete dosage compensation in both developmental stages. Median X:A expression ratios were found to be close to one in male and female larvae and pupae at various. While they note differential abundances of genes between the X chromosome and autosomes that are sex-biased in their expression, autosomal genes in this study were approximately equally expressed in both sexes, and X-linked genes were slightly female biased. However, female-bias of X-linked genes was strongest in the larvae compared to the pupae. They note that female bias of X-

linked genes in pupae could result from an absence of dosage compensation in testes (Rose *et al.*, 2015, Baker and Russell, 2011), X inactivation during male meiosis, or selection for the accumulation of female-beneficial genes on the X (Magnusson *et al*., 2012, Meiklejohn and Presgraves, 2012). Median M:F expression ratios for X-lined and autosomal loci differ significantly within all samples in this study (parental species and hybrids), but both of these ratios are near one. In all species / hybrids median M:F expression ratios are lower for X-linked genes compared to autosomal loci, suggesting, not surprisingly, an over-representation of female-biased expression genes on the X chromosome and an over-representation of male-biased genes on the autosomes. Tissue-specific dosage effects in males could also influence this pattern. These two patterns are not mutually exclusive.

The analysis of median X:A and 2:3 gene expression ratios did not find any evidence for disruption of dosage compensations in hybrids between *An. coluzzii* and *An. arabiensis,* or *An. coluzzii* and *An. quadriannulatus*, irrespective of the direction of either the cross. Median X:A gene expression ratios did not differ between males and females of these hybrids; all were near 1.0, indicating complete compensation. Spermatogenesis occurs primarily during the L4 and pupal stage in *Anopheles* mosquitoes (Clements, 1992, Krzywinska and Krzywinski, 2009). An analysis of gene expression in the *An. gambiae* male pupae germline performed by Rose *et al* (2016) suggests that, like *Drosophila* and other dipterans (Meiklejohn and Presgraves, 2012, Vicoso and Bachtrog, 2015), the X chromosome is not hyper-expressed in the testes, and thus does not show dosage compensation in this tissue (Rose *et al*., 2016, Baker and Russell, 2011). Sterile F1 male hybrids in the *An. gambiae* complex exhibit atrophied testes, malformed sperm, or a total lack of sperm due to the disruption of meiosis (Slotman *et al*., 2004). However, their somatic tissues develop normally. So, the conclusions herein regarding dosage compensation likely only pertain to male somatic tissue; dosage compensation cannot be disrupted in the male germ-line because it does not occur normally.

The rapid radiation of species in the *An. gambiae* species complex provides a unique study system for investigating fundamental questions in evolutionary biology and genetics. Despite ongoing introgression and low levels of genetic differentiation between members of the *An. gambiae* complex (Fontaine *et al*., 2015), reproductive and ecological isolation has allowed for local adaptation, behavioral divergence, and structural reorganization of their genomes (Neafsey *et al.*, 2015). This study demonstrates that despite the genetic and ecological divergence between *An. gambiae* complex member species, dosage compensation operates effectively in males to balance X-linked and autosomal gene expression levels. Additionally, this process is not impacted by hybridization, and likely does not contribute to the male hybrid sterility phenotype.

CHAPTER III

SEX-BIASED EXPRESSION & THE EFFECT OF HYBRIDIZATION ON GENE EXPRESSION IN THE *ANOPHELES GAMBIAE* COMPLEX

Introduction

Gene expression plays an essential role in converting genotypes to phenotypes, and genetic differences between species that contribute to gene expression divergence can play a role in speciation when they result in reproductive incompatibilities (Maheshwari and Barbash, 2012). Expression differences between, and within, species impact the evolution of diverse traits, for example, body color in *Drosophila* (Wittkopp *et al.,* 2003). Hybrid male sterility is a common post-zygotic reproductive barrier between species. Hybrid sterility and inviability can arise as alleles that increase fitness in pure-species genetic backgrounds fail to interact properly when brought together in hybrid backgrounds (Dobzhansky 1936). This could result from amino acid differences in interacting proteins, post-transcriptional processes effecting mRNA abundance or stability, or improper regulation of gene transcription in hybrids (gene mis-regulation) (Orr and Presgraves, 2000, Michalak and Noor, 2003).

Genes with male-biased or -specific expression, such as reproductive and accessory gland proteins, exhibit higher rates of divergence between species due to sexual selection (faster male effect) (Civetta and Singh, 1995, Singh and Kulathinal, 2000). In *Drosophila,* spermatogenesis progresses through mitosis of germ cells, meiosis, and finally the differentiation of spermatotids into mature sperm (White-Cooper *et al*., 2009). Infertility in *Drosophila* hybrid males is associated with spermiogenic (post- rather than pre-meiotic) failure associated with sperm individualization and maturation (Wu *et al.* 1992). Genes involved in this later stage of spermatogenesis have been shown be significantly mis-expressed (primarily under-expressed) in *Drosophila* hybrids in comparison to parental species (Michalak and Noor, 2003, Michalak and Noor, 2004, Moehring *et al.*, 2007, Catron and Noor, 2008, Sundararajan and Civetta, 2011, Gomes and Civetta, 2014, 2015).

Gene mis-expression in hybrids can fall into three categories: under-expression, where hybrid expression of a gene is significantly less than both parents, over-expression, where hybrid expression is significantly higher than both parents, or additive expression, where the hybrid exhibits expression that is intermediate between both parents, but still differs significantly from both. Gene mis-expression in hybrids is caused by *cis* and/or *trans* acting variants derived from parental genomes. *cis*-regulatory variants effect allele-specific transcription, and *trans*-regulatory variants effect transcription of both alleles in a diploid cell.

Cis-regulatory changes are more likely to result in additive expression than *trans*-regulatory changes within *D. melanogaster* (Lemos *et al*., 2008) and between *D. melanogaster* and *D. sechellia* (McManus *et al.*, 2010). Under- or over- expressed genes in hybrids most commonly result from compensatory *cis*- x *trans*- interacting factors that co-evolved to modulate expression in one or both parental species. When, brought together in a hybrid, the interactions between the hetero-specific *cis*- and *trans*- factors results in under- or over-expression of the target gene (Landrey *et al*., 2005). Thus, divergence of co-evolved, interacting *cis*- and *trans*- acting factors between species can have a large impact on gene mis-expression in hybrids, and contribute to the evolution of post-zygotic reproductive isolation.

In addition to its importance in the field of public health, the *An. gambiae* species complex has emerged as a model system for the study of the genetics and genomics of speciation. This is due to the recent divergence of its member species approximately 1.8 million years ago (Coluzzi *et al.* 2002), and the varying levels of reproductive isolation between various species and forms. Hybrid female F1 progeny of crosses between the six formally recognized species within the complex are fertile, and males range from semi- to completely sterile (White 1974, Hunt 1998).

I've performed a genome-wide analysis of gene expression in male and female *An. coluzzii*, *An. arabiensis*, and *An. quadriannulatus*, and bi-directional hybrids between *An. coluzzii* and *An. arabiensis*, and *An. coluzii* and *An. quadriannulatus.* By comparing these data, I've explored how hybridization results in the genome-wide mis-

expression of genes. I've focused in particular on genes that are sex-biased in parental strains because genes with reproductive functions that are mis-expressed are more likely to contribute to post-zygotic reproductive isolation. Lastly, I've explored how gene mis-exression in F1 males influences the hybrid sterility phenotype, or results from this phenotype. RNAseq data was collected from mosquitoes 12-24 hours post-pupation. Pupation in a critical developmental period during which the gonads of males and females undergo development, and spermatogenesis takes place in fertile males. Thus, at this time-point I am able to measure gene mis-expression in hybrids that contributes to fertility and impacts post-zygotic reproductive isolation.

The analysis of gene expression in this trio of species has allowed me to investigate how time-since-divergence impacts hybrid sterility and gene expression. According to a phylogenomic analysis of the X chromosome, the common ancestors of *An. coluzzii* + *An. gambiae* and *An. arabiensis* + *An. quadriannulatus* diverged 1.85 (+/- 0.47) million years ago (Ma). Next, *An. arabiensis* and *An. quadriannulatus* diverged from their common ancestor (1.28 (+/- 0.37) Ma), and *An. coluzzi* and *An. gambiae s.s.* diverged 0.54 (+/- 0.11) Ma (Fontaine *et al*. 2015, Fig 1C). In contrast to the branching order of the X, a genome-wide analysis of genetic differentiation indicates that *An. arabiensis* is more closely related to *An. coluzzii* than to *An. quadriannulatus. An. arabiensis* and *An. coluzzii* are broadly sympatric in east and central sub-Saharan Africa, and hybrids have been found at a frequency of 0.2-0.76% in field-caught specimens (Temu *et al*., 1997, Toure *et al*., 1998). Large portions of the genome have introgressed between *An. coluzzi* and *An. arabiensis* through the backcrossing of fertile F1 females to parental species (Besansky *et al.*, 2003, Fontaine *et al.*, 2015). Thus, while the X chromosome supports a branching order of *An. arabiensis* and *An. quadriannulatus* as sister species, historical and ongoing introgression between *An. arabiensis* and *An. coluzzii* has led to a lower level of genetic divergence between their genomes (Neafsey *et al*., 2015, Fontaine *et al.,* 2015).

Methods

In order to assess gene mis-expression as it relates to hybrid sterility I first analyzed sex-specific expression in each parental strain. Next, I compared F1 hybrids to both parental species of the same sex to determine which genes were mis-expressed. These were categorized as additive, under-, or over-expressed depending on their expression pattern in relation to each parental strain. In an effort to identify genes that are involved in post-zygotic reproductive isolation I created a subset of the mis-expressed genes to include only those that were sex-specific in either parental species. For example, mis-expressed genes in a F1 COLZ x QUAD male hybrid that showed male-specific expression in COLZ males and/or QUAD males. I performed a network analysis of gene interactions to identify which of these genes could potentially influence the expression of other mis-expressed genes in the dataset. For each F1 hybrid network, I performed a gene ontology analysis of the internal nodes of this network. Internal nodes were designated as those that had connections to three or more other mis-expressed genes.

Mosquito rearing, collection of F1 hybrids, RNA extraction, and RNA sequencing, sequence read quality control, and mapping were performed as described in Chapter 2 (Dosage Compensation). Raw sequence read counts per gene for each biological replicate were used as an input into the DESeq R/Bioconductor package (Anders and Huber, 2010) that was used to calculate differential expression between samples/treatments. DESeq was used to calculate the log2 fold change in gene expression between samples and calculate p-values for each comparison that are adjusted for multiple testing using the approach of Benjamini and Hockberg (1995).

A gene was considered to be sex-biased in parental strains if it has a p-adj. < 0.05 and a log2 fold change > 1 or < -1. In the male vs. female comparisons, genes that showed male-biased expression had log2 fold change > 1, and those with female-biased expression had a log2 fold change < -1. I performed a gene ontology (GO) analysis of male- and female-biased gene lists for each parental strain using the online tool g:profiler (http://biit.cs.ut.ee/gprofiler/, Reidman *et al.*, 2016), which was used to search

for the over-representation of GO terms in each dataset.

Differential expression was analyzed for each male and female F1 hybrid in DESeq by testing for genes that were mis-expressed in relation to both parental species of the same sex each to both parental species of the same sex (ex. F1 COLZ x QUAD male vs. COLZ male and QUAD male). Genes were considered mis-expressed in F1 hybrids if they had a log2 fold change > 1 or < -1 and a p-adj. < 0.05 in relation to both parents. To analyze the gene ontology of mis-expressed genes in F1 hybrids I focused on those that were sex-biased in the parental strains. The list of significantly mis-expressed genes in each hybrid was reduced to only include genes that were sex-biased (male- or female-biased) in both parental species. For example, the mis-expressed gene set of F1 COLZ x QUAD males was reduced to only include genes that overlapped the merged list of COLZ male- and female-biased and QUAD male- and female-biased genes. This analysis was performed for both male and female F1 hybrids.

Interactions between mis-expressed genes in hybrids (reduced to sex-biased parental lists) were assessed using the IntAct Molecular Interactions Database (http://www.ebi.ac.uk/intact/, Orchard *et al*., 2014), which builds a network of gene interactions based upon previously cited evidence. Genes with three or more molecular interactions were internal to the gene interaction network, and were subjected to further gene ontology analysis and exploration of gene functions using primarily VectorBase (https://www.vectorbase.org/, Giraldo-Calderon *et al.*, 2015) and FlyBase (http://flybase.org/, Gramates *et al.*, 2017).

Results

Sex-biased Expression in Parental Strains

Male-biased genes far outnumbered female-biased genes in *An. arabiensis* (ARAB), *An. coluzzi* (COLZ), and *An. quadriannulatus* (QUAD) (Table 27). The analysis of differential expression found four to six times more male- then female-biased genes depending on the species. Male and female-biased genes are non-randomly distributed among chromosome arms in each species (chi-square p-value < 0.05) with

the exception of COLZ female-biased genes (p-value = 0.13). Chromosome arm 3L had the highest proportion of its genes that were male-biased in all species (3.59% - 4.49%), while the X chromosome had the highest percentage of female-biased genes (0.69% - 1.91, Table 27). Male-biased genes comprise 2.76% - 2.87% of all genes, while only 0.43%-0.65% of genes are female-biased. As expected, female-biased genes outnumbered male-biased genes on the X chromosomes of ARAB and QUAD, but surprisingly the opposite was true for COLZ. However, this comparison could be biased because COLZ females had the lowest number of sex-biased genes (65) out of any comparison.

The low number of female-biased genes may result, only in part, from the conservative log2 fold change < -1 cutoff imposed to determine female-specific genes. MA plots in Figure 11 illustrate the lower number and lower expression intensity of female-biased compared to male-biased genes. ARAB and COLZ share roughly between 45.5% and 40.8% of female-biased genes, while males share 81%-83.5% (depending on which species is used as the numerator) (Figure 12). In COLZ and QUAD females share only 24-30% of female-biased genes, while male biased genes are shared in similar ratios to the COLZ and ARAB comparison (70.2%-88.5%). A principal component analysis confirmed that despite differences in their sex-biased gene sets, females of each parental species are more closely clustered to each other in their gene expression than they were to the male biological replicates (Figures 13, 14). Males have a higher level of variance between species in comparison to females. Additionally, biological replicates of each treatment were clustered together, indicating a close relationship between their genome-wide gene expression profiles. The x and y axis of the PCA explain 83% of the observed variance in the parental expression dataset.

I combined the sex-specific gene lists across both sexes and all species and created a heatmap showing the expression intensity of the 100 most variable genes (Figure 14). Biological replicates of each species and sex cluster together as expected (y-axis) and the colors representing expression intensity of each gene (x-axis) shows how the genes in this set are similarly expressed in each sex (Figure 14). Genes are clustered

along the x-axis according to expression similarity. This heatmap, particularly the high expression (red) block in ARAB males, further illustrates that while many sex-biased genes are shared between species, there are a fair number that are species-specific.

The gene ontology analysis did not identify a significant over-representation of any gene ontology terms in the ARAB female-biased gene list (Table 28). The QUAD female-biased list was significantly enriched for the folate biosynthesis pathway, which was also identified in COLZ. The COLZ female-biased set was also enriched for organo-nitrogen compound catabolic processes, peptide catabolic processes, and a number of exopeptidase activities (amino-, metallo-, metaloexo-, and metalloaminopeptidase activity). The genes falling into these categories are not mutually exclusive, and may overlap these similar processes.

Male-biased gene lists for all species were significantly enriched for a number of biological processes and cellular components involved in spermatogenesis and sperm motility including cillium and microtubule based movement, cillium assembly, axoneme and axonemal dynein complex assembly (the motor component of the sperm flagellum), motor activity, etc. (Table 29). There are two primary conclusions we can draw from this: (1) as expected, the male parental samples are undergoing spermatogenesis during this physiological time point (pupation) and (2) by capturing this time point we have the power to interpret the effect of hybridization on spermatogenesis in the F1 hybrids.

F1 Hybrid Mis-Expression

Mis-expressed genes accounted for 7.31% - 11.75% of all genes in F1 hybrid females (Table 30) and 8.11% - 12.15% in males (Tables 30 and 31). For both males and females, F1 QUAD x COLZ had the lowest percentage of mis-expressed genes, followed by COLZ x ARAB, COLZ x QUAD, and ARAB x COLZ. Genes that were male- or female-biased in parental strains (COLZ + ARAB = 544, COLZ + QUAD = 539, Figure 12) accounted for 0.23% - 0.42% of mis-expressed genes in F1 hybrid females, and 0.41% - 1.80 % of mis-expressed genes in F1 hybrid males (Tables 30, 31). Mis-expressed genes in hybrids were categorized as additive, over-, or under- expressed in relation to their parental strains of the same sex. In all hybrids, the majority of genes

were under- expressed, followed by over-, then additive expression (Table 32, Figures 15, 16). Males of each cross have a higher percentage of genes that are over- vs. under-expressed in comparison to females, though under- expression was still the predominant mode (Table 32). The probability of a mis-expressed gene to show additive, under-, or over- expression is not influence by its location on a chromosome arm (Figures 15, 16).

Genes found to be mis-expressed in F1 hybrids that were also in the sex-biased gene list of the parental species were assessed using the IntAct Molecular Interactions Database (http://www.ebi.ac.uk/intact/, Orchard *et al*., 2014), which builds a network of gene interactions based upon previously cited evidence. The resulting networks were comprised of internal nodes of many connections, and terminal genes at the ends of "spokes" that were influenced by an internal gene. Tables 33-40 report the results of gene function searches for the internal nodes of all F1 hybrid interaction networks. The *Drosophila melanogaster* ortholog for each gene is reported if one was identified. In many cases, the gene function is supported by evidence gained in *Drosophila* or human systems. Because these searches were informed limited to sex-biased parental genes, the results are similar to the GO term enrichment tests from parental species.

ARAB x COLZ females had mis-expressed, sex-biased genes internal to their interaction network that encode a fibrogen-like gene involved in *Drosophila* immunity, a chitin-binding protein, a cuticular protein, and a serine protease involved (Nudel) involved in embryonic dors-ventral patterning (Table 33). The analysis of the opposite direction of the cross (COLZ x ARAB females) found Nudel, the fibrogen-like gene, and also an ortholog to the *Drosophila* gene NAGLU, a lysosomal enzyme that can lead to neurological dysfunction when mutated (Table 34). Additionally COLZ x ARAB females showed significant mis-expression of TEKT3, a member of the tektin family of microtubule-associated cytoskeletal proteins. Tektins are essential for the construction of flagella and are thought to play a role in sperm competition through flagellum stability and motility.

Despite having similar numbers of mis-expressed and sex-specific mis-expressed genes to the female hybrid between COLZ and ARAB (Table 30), females resulting

from the COLZ and QUAD crosses had more nodes and links in their interaction networks (Tables 35, 36). Nudel was also mis-expressed in QUAD x COLZ females. Other mis-expressed genes include the *Drosophila* ortholog to AGAP006968 is an anion exchange protein that regulates pH and has been shown to be involved in ovarian follicle. The *Drosophila* ortholog to AGAP004787, FarO, regulates lipid storage in oenoctyes that influences tracheal waterproofing, desiccation resistance, and pheromonal communication. A gene was found that is implicated in cuticle development, and the ortholog to *Drosophila* AUST and BORR, which are involved in chromosome condensation and other processes during mitosis. The interaction network for mis-expressed genes in the opposite direction of the cross (COLZ x QUAD female, Table 36) found three proteins involved in cuticle development, ND-49, which is involved in cellular respiration, and DNA primase, which synthesizes small RNA primers for Okazaki fragments during DNA replication. In addition, three genes with *Drosophila* orthologs (MAD2, ALD/Mps1, and NMDYN-D7) were identified that are involved in microtubule polymerization and mitotic checkpoints.

Male F1 hybrids had a much larger number of genes involved in their interaction networks due to the fact that the majority of genes in the sex-biased parental gene sets are male-biased genes (Table 27, Figure 12). ARAB x COLZ males have mis-expressed genes internal to their network whose human orthologs encode sperm proteins (testis-specific-like 1, outer dense fiber of sperm tails 2 (ODF2), and ODF3), imaginal disk development, and two genes (AGAP003083 and AGAP008186) whose *Drosophila* orthologs are seminal fluid proteins (Table 37). In addition, the *Anopheles* orthologs to the *Drosophila* genes EXU and ELBA were identified. EXU is required for *Drosophila* spermatogenesis and has sex-specific splicing controlled by Tra-2. ELBA2 is involved in chromatin silencing.

In the opposite direction of the cross (COLZ x ARAB males, Table 38) human sperm proteins ODF2 and ODF3 were also identified. Two tektin proteins that are essential to the flagella were found, in addition to an additional *Drosophila* flagella associated protein (CG11449). Four *Drosophila* orthologs were identified that are

involved in flagella microtubule organization and dyneinin arm assembly inside the flagellum (CG14183, CG32392, CG7264, UQCR-C1). Lastly, the orthologs to the *Drosophila* transcription factors BIP2/TAF3 and FEST were identified. BIP2/TAF3 is involved in DNA damage response and is involved in the same pathway as LOK/CHK2 (also identified), which is a key component of double-stranded break repair during meiosis. FEST interacts with Rbp4 to direct repression of translation at cell type and stage-specific time-points during male meiosis (Table 38).

While the functions of the genes mis-expressed in QUAD x COLZ and COLZ X QUAD males are similar to those found in the COLZ and ARAB male hybrids (cell fate, morphogenesis, DNA replication, spermatogenesis), there is not much overlap in the genes between the two sets of crosses. The orthologs to two *Drosophila* Sperm-Leucyl-aminopeptidase genes (S-LAP1 and S-LAP2) were identified in the QUAD x COLZ male mis-expressed gene interaction network (Table 39). Others include the ortholog to *Drosophila* KRN and SPI, two Egfr ligands involved in cell growth, survival, and devlopmental patterning during embryogenesis and wing development. Others include CLT, involved in vitellogenesis and juvenile hormone response, FZ, involved in cell polarity, OTU, involved in oogenesis and germ cell fate, and Couch Potato (CPO), which is involved in a variety of neurological responses.

Genes mis-expressed in COLZ X QUAD F1 males had some overlap to those discussed previously (tektins, orthologs to *Drosophila* FZ, FEST, ALD/Mps1, and NMDYN-D7). Two notable differences include the ortholog to *Drosophila* Milka/Hanabi (MIL) and Nanos (NOS). MIL mutants show abnormalities in nuclear shaping and spermatid elongation. NOS is a maternally supplied *Drosophila* protein that is involved in the suppression of translation during embryogenesis. The mouse ortholog Nanos2 is primarily expressed in male germ cells where its ablation results in a complete loss of spermatogonia.

Discussion

Consistent with Haldane's Rule (Haldane 1922) genes that are male-biased in

their expression are mis-regulated at higher levels in hybrids when compared to female-biased genes (Michalak and Noor 2004, Ortiz-Barrientos *et al.* 2007) in taxa with heterogametic males, due to their higher rates of interspecific genetic divergence (Parisi *et al.* 2003) and intraspecific expression variation (Meiklejohn *et al.* 2003) when compared to female-biased genes.

In this analysis I have identified a higher proportion of genes that are male-biased than female- biased in each species. Male-biased genes have, on average, higher levels of relative expression (log2 fold change) when compared to female-biased genes (Figure 11). Twenty-seven percent of the genes in the combined set of COLZ and ARAB female-biased genes were shared between the species, and 16% were shared between COLZ and QUAD. In contrast, 70% of the male-biased genes were shared between COLZ and ARAB, and 64% were shared between COLZ and QUAD. This suggests that even if inter-specific genetic divergence in male-biased genes is higher when compared to female-biased genes, on average it has not resulted in a divergence in *which* genes are male-biased between species.

Male and female-biased genes are non-randomly distributed amongst chromosomes in all species (chi-square p-value < 0.05) with the exception of COLZ female-biased genes. This result may be influenced by the low number of genes in COLZ that show female-biased expression (Table 27). ARAB and QUAD females had the expected over-representation of female-biased genes on the X chromosome, and ARAB, COLZ, and QUAD has the highest proportion of male-biased genes on chromosome arm 3L. 3L harbors a significant sterility QTL with a relatively large effect in both a (COLZ x ARAB) x ARAB backcross (Slotman *et al.*, 2004), and a (COLZ x QUAD) x QUAD backcross (see Chapter 5). Thus, when this section of the COLZ chromosome is introgressed into an ARAB or QUAD genetic background it contributes to sterility, and has a large effect when coupled with a COLZ X chromosome. Two male-biased genes (AGAP010902, AGAP011032) are located inside this QTL and are significantly under expressed in COLZ x QUAD males when compared to parental males (Chapter 5). AGAP010902 is a cuticular protein and the *Drosophila* ortholog of

AGAP011032 is inferred to be involved in axoneme assembly, cilium movement, and cell motility, all biological functions associated with sperm proteins.

Female-biased gene sets of COLZ and QUAD are enriched for the folate biosynthesis pathway. Folates are cofactors in purine and pyrimidine ring synthesis and have been shown to effect growth rates in some insects (Blatch *et al.*, 2010). The COLZ female set was enriched for a number of aminopeptidase classes that are involved in cell de-differentiation, germ-cell niche homeostasis, and population maintenance in *Drosophila* (Lim *et al*., 2015). It is possible that these are fulfilling a similar role during female *Anopheles* to initiate ovary development and oogenesis.

Male-biased gene sets are enriched for a number of biological processes, cell classes, and molecular functions associated with cell differentiation and spermatogenesis. Of particular note are the genes that are associated with GO terms such as: axoneme, dynein arms and complexes, microtubules, cillia, and flagella. The axoneme is the central component of the sperm flagellum. It has a 9+2 structure of a central pair of microtubules surrounded by nine doublet microtubules. Doublet microtubules have an inner and outer dynein arm that serve to attach and detach to neighboring doublet microtubules, resulting in the motor activity of the flagellum. The axoneme is surrounded by a mitochondrial sheath that serves to provide ATP and power motor activity. Tekins are located near the junction points of dynein arms and neighboring microtubules (Gagon, 1995, Linck *et al.*, 2016). As such, the enrichment of gene GO terms associated with these structures and functions in the male-biased gene sets represents that the physiological process of spermatogenesis is occurring in the male pure-strain samples.

My analysis included all sex-biased genes when analyzing both male and female mis-expression, so it is difficult to know if male-biased genes were mis-expressed at a higher rate than female-biased genes. This is made more difficult by the fact that so few genes showed evidence of female-biased expression in parental strains. In all F1 hybrids, males had a higher percentage of genes that were mis-expressed in comparison to females, which is in agreement with studies in *Drosophila* (Michalak and Noor 2004,

Ortiz-Barrientos *et al.* 2007). In an analysis of hybrid vs parental gene expression involving *D. mauritiana*, *D. simulans*, and *D. sechellia*, Moehring *et al.* (2007) found that a higher proportion of hybrid genes were under vs. over-expressed when compared to parental strains. The same was true for my analysis, which found that under-expression was the predominant mode of gene mis-expression in F1 hybrids.

Gene expression analysis of sterile hybrids cannot distinguish whether gene mis-expression is linked to sterility (resulting from the sterility phenotype) or results from the interaction of incompatibilities between the parental species (ex. fast male regulatory divergence) (Gomes and Civetta, 2014). Thus, my analysis of sterile F1 hybrid males cannot interpret whether or not a gene was mis-expressed because of the sterility phenotype, or was causal to the sterility phenotype. However, there is clear evidence that gene mis-regulation is linked to sterility in F1 male *Anopheles*, and that divergence in gene regulation has resulted in gene mis-regulation in fertile F1 females.

Divergence in transcription regulation between *Anopheles* parental species has led to the mis-expression of sex-biased genes in the fertile F1 females. This includes the expression of the male-biased genes in females. TEKT3, a sperm protein, is mis-expressed in COLZ x ARAB females, and AUST, a protein specific to the chromosomal passenger complex during male meiosis (Gao *et al.*, 2008), is mis-expressed in ARAB x COLZ females (Table 35). COLZ x ARAB also showed significant mis-expression of the homolog of *Drosophila* CG8177, which causes morphological abnormalities of ovarian follicles when knocked-down (Ulmschneider *et al.*, 2016). These examples of mis-expressed, sex-biased genes, along with over a thousand other mis-expressed genes (Table 31), identified in F1 hybrid females lends support to the argument that gene mis-expression is caused by the interaction of incompatibilities between parental species in both females and males (rather than being exclusively a result of sterility).

In both species comparisons the F1 hybrid males in which *An. coluzzii* was the maternal strain (COLZ x ARAB and COLZ x QUAD) had over two times the number of mis-expressed, sex-biased genes than the opposite direction of the cross (Table 31). This could be due to deleterious trans- interactions between the COLZ X chromosome and the

interspecific autosomes. Slotman *et al.* (2004) found that COLZ x ARAB F1 males had fully arrested sperm development, whereas ARAB x COLZ F1 males had abnormal or immature sperm present. A similar pattern was found in COLZ x QUAD and QUAD x COLZ F1 males (data not reported).

A large proportion of the sex-biased, mis-expressed genes examined in F1 hybrid males are involved in cell differentiation and spermatogenesis. This observation was expected as a result of the sterility phenotype. Sterile males exhibit immotile sperm, under-developed sperm, and in extreme cases, a complete lack of sperm and atrophied testis. Thus, significant mis-expression (specifically under-expression) sex-biased genes in males may be due to the lack of, or under-development of, the reproductive tissues in which is the gene is expressed. That being said, in the COLZ x ARAB and COLZ x QUAD male gene sets a number of "higher-level" genes are mis-expressed that were central to the interaction networks. These include the significant under-expression of transcription factors LIM3 and BIP/TAF3 in COLZ x ARAB males. Additionally, nanos (NOS) is mis-expressed in COLZ x QUAD males and mutations in this gene result in a complete loss of sperm in mice. FEST is also mis-expressed in COLZ x QUAD males and plays a disproportionate role in regulating other genes during male meiosis. Further exploration into the functions of these genes may reveal that they play a large role in the sterility phenotype through the downstream regulation of other sex-biased genes.

Deleterious genetic mutations that occur during the lineage of a species are most often removed from the population by selection and genetic drift. However in some cases, compensatory mutations can mask the deleterious effects of the initial mutation. In cases where the initial mutation effects the transcription of a gene, compensatory mutations modulate transcription through cis- or trans- interactions with the initial mutation. When divergent genomes are brought together in an inter-specific hybrid, compensatory mutations may no longer compensate (for example, in trans-), and gene expression is effected. In hybrids between *An. coluzzii, An. arabiensis*, and *An. quadriannulatus,* incompatibilities between the parental genomes contribute to the mis-expression of 7.31% - 12.15% of the genes in the genome. The majority of these genes

are not sex-biased in expression. This suggests that genes with diverse suite of biological functions are affected by hybridization, which could result in a disruption of a number of behaviors and phenotypes that fall outside the scope of those that directly impact post-zygotic reproductive isolation.

While hybridization has a large effect on the sterility of F1 males, and the gene expression associated with this phenotype, only 0.23% - 1.80% of mis-expressed genes are sex-biased and likely related to reproductive processes. These results support the conclusion that divergence in transcription regulation and genetic compatibilities between the *Anopheles* species studied likely contributes to gene mis-regulation and reproductive isolation.

CHAPTER IV
HYBRID ALLELIC IMBALANCE AND DIVERGENCE IN TRANSCRIPTION REGULATION BETWEEN MEMBER SPECIES OF THE ANOPHELES GAMBIAE COMPLEX

Introduction

Genetic changes that result in gene expression differences between species have been shown to contribute to post-zygotic isolation (Maheshwari and Barbash, 2012), specifically due to divergence in cis- and trans-regulatory elements that effect transcription. Transcription is regulated through interactions between cis-acting DNA (primarily promoters and enhancers) and trans-acting RNA and proteins such as transcription factors and the RNA polymerase complex. The RNA polymerase complex acts to produce basal levels of mRNA, which can be modulated by transcription factors bound to enhancers 5' or 3' of the transcription start site. These transcription factors interact with the RNA polymerase complex through DNA looping, which is influenced by chromatin structure (Wray *et al.*, 2003).

Some genes responsible for post-zygotic isolation have DNA-binding functions (including *OdsH* and *Prdm9* in mice, and *Hmr* in *Drosophila*, and Johnson 2010), and therefore have been the focus of studies exploring the relationship between hybrid dysfunction and gene regulation. Johnson and Porter (2007) performed a simulation that demonstrated that hybrid incompatibility due to gene mis-regulation could arise when allopatric populations experience parallel directional selection acting on traits controlled by a regulatory pathway. Genes that are mis-regulated in hybrids are most often down regulated, which can be caused by transcription factor (TF) – DNA binding site divergence between lineages (Ortiz-Barrientos *et al.* 2007).

The potential for hybrid dysfunction increases with the number of loci involved in a regulatory pathway and the complexity of the binding site interactions (Johnson and Porter 2007). Highly conserved TFs will likely match their target site in a divergent lineage. Likewise, TFs that interact with highly variable sites intra-specifically will have

a higher probability of accurately binding to their respective target sites in hybrids (Ortiz-Barrientos *et al.* 2007). This is because the native genome of the TF will have adapted compensatory mechanisms to counteract low binding efficiencies in binding-site-variable species. Graze *et al.* (2012) used a novel Bayesian framework to analyze allelic imbalance (AI) in *Drosophila melanogaster* x *D. simulans* hybrid female heads. Measuring allele-specific expression of genes in hybrids and comparing expression levels to parental species can be used to identify how transcription regulation had diverged between species pairs. These authors identified significant, *D. melanogaster*-biased AI on a genome-wide scale. They also found evidence for positive selection in some AI-positive coding regions, 5' UTRs, and 5' and 3' inter-genic regions involved in gene regulation. This provides additional evidence that genetic divergence in binding sites of regulatory proteins influences gene expression in hybrids, and can contribute to hybrid dysfunction.

Cis-acting factors involve DNA regions close enough to protein coding sequences to facilitate interactions with the RNA polymerase complex, while trans-acting factors can be located elsewhere in the genome (distantly on the same chromosome, or on a different chromosome). As species diverge they can evolve changes in cis-, trans-, or a combination of the two to modulate gene transcription to ideal levels. When species hybridize, genes on homologous chromosomes share the same -trans environment that is comprised of trans-acting elements from both parental genomes, but differ in their cis-regulatory regions. This can lead to improper interactions between cis- regulatory regions and non-native trans-regulatory elements that in turn, can cause the mis-expression of a gene, or favor the transcription of one allele over another.

By comparing allele-specific expression (maternal vs. paternal) in a hybrid to gene expression in the maternal and paternal parental strains, we can gain an understanding of how parental species have diverged in transcription regulation of a gene. Parental genomes exist in a shared trans-regulatory environment in a hybrid. Therefore, if transcription regulation of a gene has diverged between two species

exclusively through trans-acting factors, we would not expect the gene to show allelic imbalance (favoring of one allele over another) in the hybrid. If allelic imbalance is observed in a hybrid, this can be inferred as divergence in cis-regulation between parental species (Cowles *et al*., 2002). If gene expression in a hybrid is outside the range observed between parental strains, and allelic imbalance is observed, transcription regulation of the gene has undergone divergence in both cis- and trans- (Wittkopp *et al*, 2004). If cis- and trans- elements have diverged between parental species (one in each), cis- x trans- interactions are observed in hybrids. If cis- and trans- factors that have diverged together in one parental species (but not the other), a pattern of cis- + trans- divergence is observed in hybrids (Coolon and Wittkopp, 2013). Within each species, cis- and trans- mutations can confer a compensatory effect to modulate transcription to ideal levels shared by both. Thus, these genes are not differentially expressed between species, but exhibit allelic imbalance in hybrids where compensatory cis- x trans- interactions break down (Takahasi *et al*., 2011).

In an effort to identify genes that have diverged in transcription regulation between members of the *An. gambiae* species complex, I've performed a genome-wide analysis of allelic imbalance in male and female bi-directional hybrids between *An. coluzzii* and *An. arabiensis*, and *An. coluzii* and *An. quadriannulatus*. I have measured expression of maternal and paternal alleles in hybrids, and have compared allele-specific expression in hybrids to maternal and paternal expression (of the same sex) in order to identify genes that have diverged in transcription regulation between parental species. I discuss how the observed divergence in transcription regulation in cis-and trans- between has been impacted by the evolutionary history of these species.

Methods

Mosquito rearing, collection of F1 hybrids, RNA extraction, and RNA sequencing, sequence read quality control, and mapping were performed as described in Chapter 2 (Dosage Compensation). In short, F1 hybrid RNAseq reads were aligned to bi-parental, cDNA pseudo-genomes that were constructed to incorporate genetic variation

observed in parental strains. F1 hybrid BAM alignment files were then processed using the program ASE-TIGAR to identify the parent of origin of each transcript. ASE-TIGAR uses a Bayesian approach to estimate transcript abundance from RNAseq data that has been aligned to diploid pseudo-genomes. ASE-TIGAR uses variation in each pseudo-genome to appropriately match each transcript to its parent of origin when fixed genetic differences exist between the parental species. Transcripts for which the parent of origin cannot be determined due to a lack of genetic differentiation between them are assigned to a parent randomly. With this approach, I quantified allele-specific expression in F1 hybrids. In contrast to previous chapters (Chapter 2: Dosage Compensation, Chapter 3: Sex-biased and Hybrid Mis-Expression), allele specific expression of maternal and paternal alleles was analyzed, rather than the combined total expression of both.

Raw sequence read counts per gene (allele for hybrids) for each biological replicate were used as an input into the DESeq R/Bioconductor package (Anders and Huber, 2010) that was used to calculate differential expression between parental gene expression (maternal vs. paternal) and expression between maternal and paternal alleles in F1 hybrids. Autosomal and X chromosome genes were analyzed separately for females and only autosomal genes were analyzed in males. DESeq was used to calculate the log2 fold change in gene expression between samples in each comparison (between alleles in hybrids or between parental strains) and to perform a chi-square test of significant differentiation between normalized expression levels. P-values are adjusted for multiple testing using the approach of Benjamini and Hockberg (1995).

A gene was considered to show allelic imbalance in a hybrid if it had a p-adj. < 0.05 in a comparison. No log2 fold change cutoff was imposed. The same criteria were used to identify differentially expressed genes between maternal and paternal parental strains. Genes were categorized as divergent in cis- only, trans- only, cis- + trans-, cis- x trans-, compensatory, or conserved. Following Coolon and Wittkop (2013), these patterns are visualized with a scatter plot comparing the ratio of maternal / paternal, parental expression (x-axis) to maternal / paternal allelic expression in hybrids (y-axis) (Figure 17). For example:

(COLZ female / QUAD female) vs.

(COLZ x QUAD F1 female COLZ allelic expression / COLZ x QUAD F1 female QUAD allelic expression)

Based upon the pattern of allelic expression in hybrids vs. parental expression, genes can be categorized as only showing allelic imbalance (compensatory), only differentially expressed between parental strains (trans- only), or showing maternal or paternal bias in hybrids and between parental strains (in the same direction) at a ratio near 1:1 (cis- only). Genes that have allelic expression in hybrids, are differentially expressed between parents, and fall between trans- only and cis- only show cis + trans divergence. Genes that have allelic expression in hybrids, are differentially expressed between parents, and fall outside of these other ranges show cis- x trans- divergence. Genes that do not exhibit allelic imbalance and are not differentially expressed between parents are conserved in their transcription between parental species.

Results

Comparison Between *An. coluzzi* and *An. arabiensis*

The number of genes in each transcription category was not influenced by the direction of the cross in ARAB x COLZ and COLZ x ARAB female hybrids (Table 41). However, the direction of the cross had a significant effect on the number of genes that were categorized as compensatory in males (chi-square p-value = 0.00, Table 42). Males and females of these crosses had 72% - 76% of all autosomal genes classified as conserved (Tables 41,42). The next highest category for females was trans- only (~12% in both crosses), followed by compensatory (~11%) (Table 41). This order is switched for males, which had 13-14% of their genes categorized as compensatory, followed by ~6% trans- only (Table 42). The number of autosomal genes that are conserved, compensatory, trans- only, and cis- x trans-acting differs significantly between males and females of the ARAB x COLZ cross and the COLZ x ARAB cross (chi-square p-value < 0.05, Table 45).

The distribution of allelic vs. parental expression in these crosses can be

visualized in Figures 18 (female) and 20 (male). Conserved genes cluster in the center because they have a 1:1 allelic ratio in hybrids and a 1:1 ratio between parental strains (not significantly differentiated). Figure 17 can be used as a key to understand which regions of the scatter plot represent each transcription category. Genes that are significantly differentiated in their expression between alleles in hybrids and/or parental strains are plotted in red. The higher proportion of genes classified as compensatory in males vs. females can be seen as the genes grouping around the vertical red line in Figure 20, as compared to the females in Figure 18.

The number of genes in each transcription category does not differ significantly between either direction of the cross for X chromosome genes in females (Table 43). The distribution of the genes in Figure 22, which compares female X chromosome gene expression between the crosses, is essentially rotated 180 degrees top to bottom, indicating that the expression mode of the majority of genes is maintained irrespective of the direction of the cross. After correcting for total number of genes on the autosomes and X chromosomes, respectively, the proportion of female autosomal genes and X chromosome genes in each category does not differ significantly (Table 44).

Comparison Between *An. coluzzi* and *An. quadriannulatus*

Hybrids between *An. coluzzi* and *An. quadriannulatus* have very different patterns of expression when compared to *An. coluzzi* and *An. arabiensis* crosses. Males and females of COLZ x QUAD and QUAD x COLZ crosses differ significantly in the number of genes in every expression category (Table 45). QUAD x COLZ females have significantly fewer genes classified as compensatory compared to COLZ x QUAD females (Table 41). The majority of genes in females of these crosses are categorized as trans-acting only (45-46%), which can be visualized as the cone-like protrusions pointing from the center of the distributions in Figure 19. These genes are significantly up regulated in the parental COLZ females compared to QUAD females. Maternal bias results in a positive log2 fold change, and paternal bias is negative. Only 33-34% of genes in females are categorized as conserved in females of the COLZ and QUAD crosses compared to 72% in the COLZ and ARAB cross females (Table 41).

The number of female X chromosome genes in each transcription category did not differ significantly between directions of COLZ and QUAD the crosses (Table 43). Within a cross, the ratio of genes in each category did not differ between the X chromosome and the autosomes (Table 44). The distribution of genes looks similar when comparing parental to allelic expression of X chromosome (Figure 23) and autosomal genes (Figure 19). The large number of genes that are effected by trans- divergence are seen as a horizontal cone pointing in the direction of COLZ-biased expression.

Males of the COLZ x QUAD and QUAD x COLZ crosses do not differ significantly in the number of genes in each transcription category (Table 42). The highest proportion of genes falls into the conserved category (59%), followed by trans- (~20%) and compensatory (13%). Similar to females of these crosses, the majority of trans- diverged genes are COLZ-biased in their expression (Figure 21). Males of all crosses have a similar number of genes classified as compensatory. However, there is a ~15% shift of genes between the conserved category of the COLZ and ARAB crosses to the trans- only category of the COLZ and QUAD crosses (Table 42). This is evident when comparing Figure 20 (COLZ and ARAB hybrid males) and Figure 22 (COLZ and QUAD hybrid males). Figure 22 has a larger number of genes that show significant allelic imbalance and/or divergence between parental strains (red points), and these distributions are pulled further away from center, indicating that more genes have a higher degree of log2 fold change between maternal and paternal alleles in hybrids (more extreme allelic imbalance), and parental strains (more pronounced divergence in expression between species). The same is true when comparing females between the COLZ and ARAB versus COLZ and QUAD. Over double the number of genes show divergence in cis-, trans-, or some combination of the two.

Discussion

Previous studies of *Drosophila* inter-specific hybrids identified highly variable percentages of genes that exhibit cis- (12%-88%) and trans- (16%-78%) divergence, or the combined effect of the two (4%-65%) (Coolon and Wittkopp, 2013). In the largest of

these studies, McManus *et al.* (2010) used RNAseq to measure allelic expression of 9,966 genes in *D. melanogaster* x *D. sechellia* F1 female hybrids and parental strains. These species diverged ~1.2 million years ago (Cutter, 2008). The authors identified 5,042 (51%) genes that had diverged in cis-, 6,546 (66%) that had diverged in trans-, and 3,473 (35%) that had diverged in some combination of both. *An. coluzzii, An. arabiensis* and *An. quadriannulatus* diverged from their common ancestor 1.85 (+/- 0.47) million years ago, and subsequent introgressions in both directions between *An. coluzzi* and *An. arabiensis* has resulted in large portions of the genome (ex. the *2La* chromosomal inversion) being exchanged (Fontaine *et al*. 2015, Fig 1C). The analyses of female hybrids between these species found that ~5% of genes between *An. coluzzi* and *An. arabiensis* have diverged in cis-, ~18% in trans, and ~19% in one or both. In contrast, ~10% of genes between *An. coluzzi* and *An. quadriannulatus* have diverged in cis-, ~56% in trans, and ~57% in one or both.

During my analysis I was conservative about categorizing genes as cis - only, which may have led to the smaller proportion of genes seen in this category in comparison to the *Drosophila* study. I categorized genes as cis-only if they (1) showed allelic imbalance in hybrids, (2) were differentially expressed between parental species, and (3) had a maternal / paternal hybrid vs. maternal / paternal parental expression ratio between 0.95 and 1.05 (very close to the diagonal, cis- only lines in the plots) (Figure 17). That being said, genes above or below this 0.95-1.0 ratio fall into the cis- x trans- and cis- + trans- categories. So, they are still included in the "one or both" cis- and trans- percentages. The percentage of genes that have diverged in cis- and trans- , or both, is comparable between the the *An. coluzzi* - *An. quadriannulatus* study I have performed, and the results of the McManus *et al*., (2010) *Drosophila* study. Another important note is that the *Drosophila* study did not separate the X chromosome from the autosomes in their analyses. While I did not find significant differences in the proportion of genes that fell into each transcription category between the X chromosome and the autosomes in my analysis of females I don't know if the same is true for *Drosophila.*

The higher number of genes that are classified as conserved in the *An. coluzzi* and *An. arabiensis* comparisons, versus having diverged in cis- or trans-, is likely due to the broad ranges in which these species are sympatric and their ongoing introgression (Fontaine *et al.*, 2015). Introgression of adaptive alleles between these species will result in shared cis- acting variants that effect transcription locally, and will also select against trans-acting factors that are not compatible with a foreign allele that provides a greater selection coefficient.

In all crosses / species comparisons, males had significantly higher numbers of genes categorized as conserved and compensatory, and significantly fewer genes classified as -trans. It is important to note the trans- regulatory environment differs between males and females of the same direction of a cross. Maternal, trans-acting factors that originate from the X chromosome and affect autosomal expression are hemizygous in males. In addition, there are more male-biased than female-biased genes (Chapter 3), and the Y chromosome is present. Trans-only divergence is characterized by genes that do not exhibit allelic imbalance but are differentially expressed between parental strains. To verify that the higher number of genes categorized as trans - divergent in females is driven by higher levels of gene expression divergence between females of the parental strains in comparison to males, I analyzed differential gene expression between COLZ and ARAB females, COLZ and QUAD females, and the male datasets of each of these comparisons. In the COLZ-ARAB comparison, females had 2,335 significantly differentiated genes (p-adj < 0.05), while males had 1,467. In the COLZ-QUAD comparison females had 7,799 significantly differentiated genes (p-adj < 0.05), while males had 3,782 (Figure 24).

The larger number of significantly differentiated genes between COLZ and QUAD females and males compared to COLZ and ARAB females and males is interesting because many of these fall into the trans- only divergence category. *An. quadriannulatus* has a more limited distribution in southern Africa (White, 1974, Coetzee and le Sueur, 2000) in comparison to *An. coluzzi* and *An. arabiensis*, which are broadly sympatric across central and West Africa. The geographical isolation of *An.*

quadriannulatus, and substantially lower levels of introgression with *An. coluzzi*, has lead to higher levels of genetic divergence between *An. coluzzi* and *An. quadriannulatus* than between *An. coluzzi* and *An. arabiensis* (Fontaine *et al.*, 2015). This as led to a higher proportion of genes being differentially expressed between the two comparisons (Figure 24), in part at least, to a larger number of trans-acting elements.

A large number of the genes that have diverged in trans- in COLZ x QUAD and QUAD x COLZ females show COLZ-biased expression between the parental strains (the "cones" in Figures 19 and 24). This might lead one to think that trans-acting elements have diverged at a faster rate in COLZ than QUAD, or are acting specifically in COLZ. This is not correct. For all genes that are divergent in their transcription regulation in cis-, trans-, or some combination of the two, we only know that they differ between the species, not the species in which transcription regulation diverged. This is because we do not know if this divergence has led to the up- or down-regulation of a gene in one species or another, only that they now differ significantly. For example: if a gene had the same expression in the common ancestor of COLZ and QUAD, and we observe COLZ-biased expression, we do not know if cis- and/or trans- regulatory divergence has led to an up-regulation of that gene in COLZ or a down-regulation in QUAD. Both would result in a log2 fold change in the favor of COLZ. In the future, an analysis of genetic divergence in 5' and 3' UTRs surrounding genes that have diverged in cis- could shed light onto the direction of the effect.

Proper transcription regulation plays an essential role in development, interaction with the environment, behavior, and requires a multitude of interactions between proteins, genes, and feed-back on the physiological state and needs of cells and tissues. Genetic changes that disrupt proper regulation of gene expression are quickly purged from populations by selection. Those that occur and modify transcription to produce a selective benefit can quickly become fixed in a population and result in divergence in transcription regulation between species. These changes can be compensatory, or act in concert with other changes in a way that is not easily measureable in pure species. When brought together in a hybrid, the differences in transcription regulation between species

become apparent. In this study I have found that divergence in transcription regulation is abundant between species in the *An. gambiae* species complex. Ongoing introgression between *An. coluzzi* and *An. arabiensis* has led to lower levels of transcription divergence when compared to *An. coluzzi* and *An. quadriannulatus,* which is evidenced by a higher proportion of genes that differ in their trans- regulation.

CHAPTER V

THE GENETICS OF MALE STERILITY IN HYBRIDS BETWEEN *ANOPHELES COLUZZII AND AN. QUADRIANNULATUS*

Introduction

The *Anopheles gambiae* Species Complex

The *Anopheles* (*An.*) *gambiae* complex is comprised of seven largely morphologically indistinguishable species. All members of the *An. gambiae* complex are competent vectors of human malaria parasites, though they differ in their host specificity, ranging from highly anthropophillic (*An. gambiae*) (Takken and Knols 1990, Dekker and Takken 1998, Pates *et al.* 2001b) to almost entirely zoophilic (*An. quadriannulatus* A) (White 1974, Gibson 1996, Dekker and Takken 1998, but see Pates *et al.* 2001a). It is estimated that in 2010, approximately 216 million cases of malaria occurred globally. Of over 655,000 deaths attributed malaria in 2010, 86% occurred in children under the age of five (World Health Organization 2011). In areas of stable malaria transmission, the risk of mortality drops drastically once children reach the age of five due to acquired immunity resulting from repeated malaria infections. Due to their high preference for human blood meals, close association with human domiciles and agriculture, and their tendency to enter houses to feed, *An. gambiae, An. coluzzi,* and *An. arabiensis* are the primary vectors of human malaria in sub-Sahara Africa where approximately 90% of all deaths attributed to malaria occur (World Health Organization 2011).

The members of the *An. gambiae* species complex were thought to be a single species until experimental crosses between five 'strains' revealed the presence of hybrid sterility in F1 males **(**Davidson 1962, White 1974, Hunt *et al.* 1998). South African and Ethiopian populations of *An. quadriannulatus* were discovered to be two separate species (A and B, respectively) by Hunt *et al.* (1998), and *An. quadriannulatus* B was subsequently re-named *An. amharicus* (Coetzee *et al.* 2013). Further subdivision exists with in the complex as genetic differentiation within *An. gambiae s.s.* (hereafter *An.*

gambiae) led to the description of five "chromosomal forms" (Touré *et al.* 1983, Coluzzi *et al.* 1985, Petrarca *et al.* 1987) which are now thought to be local adaptations to various ecotypes.

Investigations of genetic differentiation between *An. gambiae* chromosomal forms revealed the presence of two molecular forms that are characterized by fixed differences in the ribosomal DNA (Favia *et al.* 1997). These were known as the M and S molecular forms and were considered incipient species within *An. gambiae* (della Torre *et al.* 2001, 2005), until *An. gambiae* M-form was recently elevated to species-level and renamed *An. coluzzii* (Coetzee et al 2013). Slotman *et al.* (2007b) discovered further genetic subdivision between M form *An. gambiae* (*An. coluzzii*) from Mali and Cameroon, which was associated with "Mopti" and "Forest" chromosomal forms. Despite the presence of fixed chromosomal inversions and molecular forms within *An. gambiae*, low levels of population genetic structure have been found across large geographic distances within *An. gambiae* molecular forms (Lehman *et al.* 2003). A similar observation was found in its sister species, *An. arabiensis* (Besansky *et al.* 1997; Donnelly and Townson 2000), but not the brackish water-breeding species *An. melas,* which is comprised of three sub-groups that exhibit high levels of genetic differentiation throughout their genomes (Deitz *et al.,* 2012, Deitz *et al.*, 2016, Chapter 1).

From a practical standpoint, it is important to know which genomic regions introgress between species because there is potential for genes that confer insecticide resistance to pass from one species to another (e.g. the movement of *knock down resistance* (*kdr*) between chromosomal forms of *Anopheles gambiae* (Weill *et al.* 2000)). Indoor residual spraying (IRS) of insecticides and the distribution of long lasting insecticide-treated bed nets (LLINs) remain the primary methods used to limit contact between humans and malaria vectors, thus the spread of insecticide resistant alleles between species can severely undermine malaria control efforts.

Ultimately, any form of speciation must lead to the formation of pre- or post-zygotic reproductive isolation. The barriers of pre- and post-zygotic isolation can be broken down by gene flow between populations (or putative species) upon secondary

contact. The persistence of these barriers to gene flow depends on their propensity to prevent mating (pre-zygotic) or the production of viable offspring (post-zygotic) (Ramsey *et al.* 2003).

Pre-zygotic Isolation

Pre-zygotic, or pre-mating, isolation occurs either through selection against migrants, whereby a migrant's fitness is lowered in its new environment, or through assortative mating between individuals belonging to two different evolutionary lineages (Schlüter and Conte 2009). In the *An. gambiae* complex, strong pre-mating isolation exists between sympatric species. For example, low numbers of F1 *An. gambiae/An. arabiensis* hybrids are found in nature (0.02%-0.76%, Temu *et al.* 1997, Touré *et al.* 1998). However, low levels of hybridization are enough for extensive introgression between the two species (Besansky *et al.* 2003, Slotman *et al.* 2007a), due to the production of fully fertile F1 female hybrid offspring (Slotman *et al.* 2005a) that can backcross to either parental species.

Milligan *et al.* (1993) were able to distinguish between 90% of sympatric *An. gambiae* and *An. arabiensis* from Banambani, Mali based upon gas chromatography of four cuticular hydrocarbons. However, they were unable distinguish between these species from another sympatric population in Moribabougou, Mali. This suggests that cuticular hydrocarbons may play a role, albeit not an exclusive one, in prezygotic isolation between *An. gambiae* complex species. There is no evidence for post-zygotic reproductive isolation between M and S molecular forms of *An. gambiae* (*An. coluzzii* and *An. gambiae s.s.,* respectively) (Diabate *et al.* 2007), however, they mate assortatively (Aboagye-Antwi *et al.*, 2015). Possible mechanisms for pre-zygotic isolation between M and S molecular forms include mate choice for inter-form wing morphology (Sanford *et al.* 2011), wing-beat frequencies or "flight-tone" (Pennetier *et al.* 2010), and mating swarm segregation between forms.

Mating swarms of *An. gambiae* M and S molecular forms are monotypic. Molecular forms mate separately over natural landmarks (M) or bare ground (S) (Diabate *et al.* 2009) in areas where gene flow between forms is low or absent (Diabate

et al. 2011). In Burkina Faso the majority of swarms are segregated by molecular form, however, mixed M and S form swarms are have been observed (15.3%, Diabate *et al.* 2006, ≈20%, Dabire *et al.* 2013). However, within mixed-form swarms assortative mating occurs and hybrids are rare (Dabire *et al.* 2013, della Torre *et al.* 2001). Pre-mating isolation breaks down in sympatric populations of M and S for *An. gambiae* Guinea Bissau. In their 2011 study, Marsden *et al.* collected 323 *An. gambiae* mosquitoes and found that this population was comprised of 13.9% M form, 52.6% S form, and 33.4% M/S hybrid mosquitoes and F1 M/S hybrids backcross to S form *An. gambiae* at a higher rate expected by chance. Mating preference between M and S molecular forms of *An. gambiae* has been shown to be linked to the X chromosome marker that is divergent between, and diagnostic of, these species (the so-called "island of speciation", Aboagye-Antwi *et al.*, 2015). While most studies addressing pre-zygotic isolation within the *An. gambiae* complex have focused on the M and S molecular forms of *An. gambiae,* mechanisms that mediate pre-zygotic isolation between these forms may also play a role in pre-mating isolation throughout the *An. gambiae* complex.

Post-zygotic Reproductive Isolation

The evolution of post-zygotic isolation mechanisms is positively correlated with the time of divergence between two putative species, and is dependent on a number of variables including: the number of mutations separating two allopatric lineages at any given time, the number of incompatibilities per mutation, and the fitness of each incompatibility (Orr and Turelli 2001). The study of speciation genetics and genomics has focused on hybrid sterility and inviability, which are forms of hybrid dysfunction (Coyne and Orr 2004, Dobzhansky 1937), because the evolution of hybrid dysfunction factors represents the first step in the formation of post-zygotic reproductive isolation between diverging populations.

Post-zygotic isolation evolves as Dobzhansky-Müller incompatibilities (DMI) accumulate between diverging populations (Dobzhansky 1936, Müller 1940). DMI arise when interacting (epistatic) or homologous genes or loci evolve in allopatry. Upon secondary contact between divergent populations, interacting or homologous genes/loci

on parental chromosomes are unable to function effectively in a hybrid. The resulting gene products in hybrids can have unpredictable, deleterious epistatic interactions (Maheshwari and Barbash 2011), changes in gene expression can result in allelic imbalance in coding and regulatory regions due to incompatible *cis*- or *trans*-acting factors (Graze *et al.* 2012), and segregation distortion of parental genes can skew sex ratios of progeny (Orr and Irving 2005, Phadnis and Orr 2009).

DMI resulting in post-zygotic reproductive isolation can evolve as a result of various mechanisms in either an ecological speciation (divergent evolution due to selection pressures imposed by novel environments) or mutation-order speciation scenario (fixation of different, incompatible mutations in allopatric populations experiencing the same environmental conditions) (Mani and Clarke 1990, Nosil and Flaxman 2011). These mechanisms include mutation-driven co-evolution (e.g. copper tolerance in *Mimulus guttatus*, MacNair 1983, MacNair and Christie 1983), gene duplication (e.g. the *OdsH* gene in *Drosophila simulans* x *D. mauritiania* backcross males, Ting *et al.* 2004, Sun *et al.* 2004), gene transposition (e.g. the *JYAlpha* gene in D. *melanogaster* x *D. simulans* backcross males, Masley *et al.* 2006), or a molecular arms race (e.g. divergent NB-LRR alleles in *Arabidopsis thaliana* strains, Bomblies *et al.* 2007, Bomblies and Weigal 2007, Bakker *et al.* 2006).

F1 hybrid female progeny of crosses between the eight formally recognized species in the *An. gambiae* complex are fertile, and, with the exception of the *An. gambiae* - *An. coluzzi* species pair, males range from semi- to completely sterile (White 1974, Hunt 1998). Previous work has identified the regions in the *An. coluzzii* (C) genome responsible for male and female sterility when introgressed into an *An. arabiensis* (A) genetic background, and vice versa, by conducting QTL analyses of sterility in CAxA and CAxC backcrosses (Slotman *et al.* 2004, 2005b). Genes located on the X chromosome have a large effect on the hybrid sterility phenotype in both directions of the cross. Additionally it was found that inviability factors exist between the *An. coluzzii* X chromosome and both *An. arabiensis* autosomes. Recombination does not occur between the X chromosomes of these species because it is suppressed by the

Xag inversion, which is fixed within *An. coluzzii* but absent in *An. arabiensis.* The *Xag* inversion comprises ≈75% of the *An. coluzzii* X chromosome. This limits the ability to map genes to the X between *An. gambiae An. coluzzi,* which share the *Xag* inversion, and any other species in the complex.

Significant male sterility QTL were found on the 2nd and 3rd chromosomes in the CAxC and CAxA backcrosses (Slotman *et al.* 2004). These sterility QTL are located on different regions of the autosomes in each backcross, which is consistent with the Dobzhansky-Müller model (Dobzhansky 1936, Müller 1940), and previous results from *Drosophila* (Wu and Beckenback 1983, Orr and Coyne 1989). A QTL on the 2nd chromosome of *An. arabiensis* causes hybrid sterility when introgressed into an *An. coluzzii* genomic background. This QTL is located within the 2*La* inversion, which is shared by *An. arabiensis* and *An. coluzzii* (Coluzzi *et al.* 2002). This finding is interesting because genes involved in hybrid sterility and inviability are thought to be present at a higher rate within non-shared inversions (Rieseberg *et al.* 1999, Noor *et al* 2001). However, Slotman *et al.* (2004) describe that the 2*La* inversion introgressed into *An. coluzzii* from *An. arabiensis* during their divergence, and it is possible that the sterility factors evolved after this. Their analyses also indicate that large regions of the *An. coluzzii* and *An. arabiensis* autosomes have little to no effect on hybrid sterility. In the CAxA cross, inviability is caused by an incompatibility between the *An. coluzzii* X chromosome and at least one locus on each autosome of *An. arabiensis.* Complete sterility is achieved through the inheritance of at least three to four autosomal sterility loci (Slotman *et al.* 2004).

Female hybrid sterility QTL are present on the third chromosomes of both *An. coluzzii* and *An. arabiensis*, though again, these are located on different regions of the chromosome in each species (Slotman *et al.* 2005b). A QTL on the second chromosome *An. coluzzii* causes female hybrid sterility when it is introgressed into an *An. arabiensis* genomic background. However, the female sterility phenotype is affected primarily by the interaction of a foreign X chromosome with the two autosomal QTL. Approximately 75% of observed phenotypic variance is due to the X chromosome, which only

represents 8.8% of the genome (Holt *et al.* 2002). Due to the lack of recombination between X chromosomes, the number of loci on the X chromosome that effect the hybrid sterility phenotype is unknown. A comparison of the data between the male (Slotman *et al.* 2004) and female (Slotman *et al.* 2005b) *An. coluzzii* x *An. arabiensis* hybrid sterility data shows that there are a greater number of male than female hybrid sterility QTL, indicating that male sterility factors evolve faster than female sterility factors. This corroborates evidence of faster male evolution in *Aedes* mosquitoes (Presgraves and Orr 1998) and higher numbers of male versus female sterility factors in *Drosophila* (Hollocher *et al.* 1996, True *et al. 1996*, Tao and Hartl 2003).

In order to determine how post-zygotic reproductive isolation evolves in the context of the *An. gambiae* species complex, I have performed a QTL analysis of male hybrid sterility in a (*An. coluzzi* x *An. quadriannulatus*) x *An. quadriannulatus* backcross. I have identify regions of the *An. coluzzi* genome that contribute to male hybrid sterility when introgressed into an *An. quadriannulatus* genetic background (CQxQ). I discuss how these QTL relate to those identified in the *An. coluzzi* x *An. arabiensis* cross Slotman *et al.* (2004), and have identified a list of male-biased genes found inside sterility QTL shared by these crosses that are mis-expressed in F1 hybrid males and may contribute to the male hybrid sterility phenotype.

Methods

Backcross and Phenotype Scoring

I performed a backcross between the SUA2La strain of *An. coluzzii* (C) and the SANGUA strain of *An. quadriannulatus* (Q) to identify regions of the C genome that exhibit DMI with the Q genome, resulting in F1 male sterility. The *An. coluzii* strain was isolated by Mario Coluzzi from Suakoko, Libera and the *An. quadriannulatus* strain was isolated by Willem Takken in Sangwe, Zimbabwe. Both strains have been lab reared for hundreds of generations. I used a standard backcross scheme that has been used in several studies to investigate the genetic basis of hybrid sterility and inviability beginning with Dobzhansky (1936) (Figure 25, left panel). F1 CQ (C females x Q males)

hybrids were reared and backcrossed to Q males (CQxQ). Backcross males have a recombinant 2nd and 3rd chromosome, an *An. quadriannulatus* Y, 2nd, and 3rd chromosome, and a hemizygous X chromosome inherited from C or Q. We did not anticipate any recombination on the X because of the *Xag* inversion in *An. coluzzii*.

Backcross males range from fully fertile to completely sterile. The testes of adult male CQxQ mosquitoes were dissected using a dissecting microscope and transferred to a compound microscope. Testes were squashed to release sperm (if present) into solution. Based upon the abundance and morphology of the sperm, and the morphology of the testes, each mosquito was assigned a phenotype score based on a one to seven scale:

1. Normal testes and sperm
2. Slightly abnormal sperm
3. 50/50 normal and abnormal sperm
4. Mostly abnormal sperm
5. Entirely abnormal sperm
6. No sperm present
7. No testes present

After being phenotyped, mosquitoes were reserved in 100% ethanol.

DNA Sequencing

DNA extractions were performed on a Qiagen Biosprint DNA extraction machine (Qiagen Inc, Valencia, CA). Extracted DNA was suspended in 200µl elution buffer and stored at -20°C. Each mosquito was genotyped using the restriction enzyme-based genotype-by-sequencing approach according to the multiplexed shotgun genotyping (MSG) approach published by Andolfatto *et al.* (2010) (Figure 25, right panel). This protocol has been used successfully to identify QTL associated with salinity tolerance in *An. coluzzii* x *An. merus* hybrids (Smith *et al.* 2015), map a dominant marker within 104,069 bp of its true location in the *D. simulans* genome (Andolfatto *et al.* 2010), and identify QTL associated with pupation site preference in *D. sechellia* and *D. simulans* (Erezyilmaz and Stern 2013).

In short, 10 ng of genomic DNA of each mosquito was digested with the restriction enzyme MseI, which cuts the *An. gambiae* genome on average every 183.2

bp. Samples were digested for three hours at 37˚C, followed by a 20 minute, 65˚C enzyme deactivation step. Non-ligated bar-coded adapters were removed using a isopropanol precipitation, and bar-coded samples with unique adapters were pooled (96 per plate). Pooled samples were precipitated via centrifugation and re-suspended in TE buffer. Next, DNA was bead-purified using an Agencourt AMPure PCR purification kit. Bar-coded, purified DNA was size-selected to a range of 250-400 bp using a Pippen Prep and then amplified using a Phusion High-Fidelity PCR kit. Samples were amplified using a common primer, and an indexed, library-specific primer (one per plate) that also incorporates the Illumina flow-cell adapter. Thus, individuals were identifiable to a library prep by their index and to the individual level by their bar-coded adapter. Library amplification (PCR) was performed at a range of 12-18 cycles to identify the minimum number necessary to amplify the library to a target of 10-30 ng/uL (post clcean up) in an effort to mitigate the incorporation of PCR errors. Amplified libraries were bead-purified using an Agencourt AMPure PCR purification kit. Library size distribution, integrity, and DNA concentration was confirmed by running each library on a NanoDrop and BioAnalyzer prior to sequencing.

This project was performed concurrently with an analysis of CQxQ female hybrid sterility (not reported here). In total, 841 CQxQ males and 482 females where phenotyped and sequenced. In addition, library prep and sequencing was performed on 192 *An. coluzzi*, and 192 *An. quadriannulatus* individuals in order to identify fixed genetic difference between parental strains that would be diagnostic of the origin of genomic regions introgressed from *An. coluzzii* into *An. quadriannulatus* in the backcross. In total, the MSG library prep was performed on 1,682 mosquitoes across twenty 96-well DNA plates (not all were full). Hybrid libraries were sequenced across three lanes of Illumina HiSeq 2500, two rapid runs using 125 bp single-end chemistry, and one high throughput run using 125 bp single-end chemistry. Parental libraries were sequenced on a single lane of Illumina HiSeq 2500, with 125 bp paired-end, high throughput chemistry.

Sequence Mapping and Genotyping

Raw reads were de-multiplexed by their library index by the core facility. I de-multiplexed raw reads to the individual level using the process_radtags program of the STACKS package (Catchen *et al.*, 2013). I used the FastQC package version 0.11.4 (Babraham Bioinformatics) to visualize DNA sequencing run and read quality. I used Trimmomatic version 0.30 (Bolger *et al.,* 2014) to trim reads by quality by simultaneously soft-clipping the reads from both 5' and 3' ends to an average phred quality score of 20, with no single bp in a four bp window below a phred quality score of 20. Only reads ≥ 50 bp were retained for alignment and subsequent analyses.

Prior to genotyping male hybrids using the suite of tools in the STACKS packages, I called and screened variants in parental libraries to ensure that STACKS was provided high quality parental alignments with which to call highly supported fixed differences between the species. Reads of parental *An. coluzzi* and *An. quadriannulatus* individuals were aligned to the *An. gambiae* AgamP4.4 genome (Holt *et al.,* 2002) because this represents the most well annotated genome out of the all the species in the *An. gambiae* complex. Additionally, it is the only one that is assembled into chromosomes (Neafsey *et al.*, 2015). Alignments were performed using the program Stampy (Lunter and Goodson, 2011) with a prior substitution rate = 0.02. Stampy is designed to map DNA sequencing reads to a divergent reference genome and has been previously used for this purpose in the *An. gambiae* species complex (Smith *et al.,* 2015). I sorted SAM alignment files, converted them to BAM format, and filtered them to a minimum mapping quality score (MAPQ) of 50 (very high) using SAMtools version 0.1.19 (Li *et al.,* 2002). Next, I identified variants between parental individuals and the reference genome with SAMtools mpileup and BCFtools call, with a prior substitution rate of 0.02. Next, I filtered variants using the SAMtools vcfutils.pl varFilter script, imposing depth threshold of 5 reads to support each variant in an individual. This script was also used to remove variants that occur within 3 bp of a gapped alignment, or more than two alternate bases. Variant call files were merged between the 192 individuals of each parental species, while imposing a filter requiring each variant to be supported by

5% of the individuals within a species. Finally, the filtered alignment files for all parental individuals of each species were merged together to create one "super-parent" alignment file for both *An. coluzzi* and *An. quadriannulatus*. These super-parent alignments were then filtered to only include alignments that overlap with the highly supported variants identified in that species (minimum 5 reads supporting a variant in 10 individuals) using the intersectBED function of BEDtools (Quinlan and Hall, 2010). These parental alignment files were used as the input to the STACKS suite of packages to call fixed SNPs between parental species.

Backcross males were aligned to the *An. gambiae* AgamP4.4 genome using Stampy with a prior substitution rate = 0.02. BAM alignment files were filtered to only include alignments with a MAPQ > 20. The STACKS program pstacks was used to identify genomic alignments *An. coluzzii* and *An. quadriannulatus* that harbor fixed SNPs with a read depth greater than 50x. These loci were merged into a catalog using cstacks. The resulting catalog contains overlapping "stacks" (alignments) from both species that contain fixed SNPs between them. Pstacks was run on CQxQ hybrids to identify fixed and polymorphic SNPs that were supported by at least five reads. Next, the SNPs from each CQxQ individual were queried against the parental catalog to find overlapping alignments in which to call SNPs in the hybrid. The STACKS program "genotypes" was used to collate genotypes from all CQxQ individuals that matched a locus in the parental catalog and write the genotypes of each individual / locus to an output file. At this point I required that locus be genotyped in ten individuals to be written to the output file. However, the number of loci per individual, and individuals per loci, was filtered further in subsequent analysis.

QTL Mapping

I used R/qtl to further filter loci and individuals from the analysis, create a genetic map, and perform QTL mapping, along with R/qtlbim, which was used to perform Bayesian Interval Mapping (Yandell *et al.*, 2007). I used R/qtl to remove markers from the dataset that were not genotyped in at least 120 individuals, and then removed individuals from the analysis that were not genotyped at a minimum of eight

markers. The number of markers was further filtered from the dataset by removing those that supported very large expansions in the genetic map or had high LOD scores for probable genotyping errors. CQxQ male genotypes were analyzed to identify genotyped markers that supported very tight double crossovers, or double crossovers that were not supported by at least three markers. This was a highly iterative process. Each time markers with high error probabilities were removed, the genetic map contracted, and thus the window size for analyzing "tight" double cross-overs was reduced. The integrity of the map was analyzed by calculating the recombination frequencies of markers, and the respective LOD scores for each estimate after each iteration using a likelihood ratio test. Marker order was forced on the initial genetic map by in the order in which the markers occur in the genome. After the final iteration of removing low probability genotypes and markers that contributed disproportionately to the expansion of the genetic map, I performed a de-novo analysis of marker order (not informed by marker order in the genome). This analysis found that the given, genomic order of markers had the highest probability.

Through this process, and in the final dataset of markers and individual genotypes, the genetic map was estimated using the Kosambi map function (Kosambi, 1943). The final genetic map was used for three types of QTL analysis: simple interval mapping (SIM), composite interval mapping (CIM), and Bayesian interval mapping (BIM). Genotypes probabilities were estimated for all individuals using a step width of zero (at every marker location). SIM was performed to first understand the general location of QTL and to assess significance of the LOD scores associated with these peaks. SIM was performed using a Haley-Knott regression (Haley-Knott). I performed 10,000 permutations of this analysis to generate a LOD score distribution with which to calculate a 5% confidence threshold. CIM randomly selects markers throughout the genome to use as co-factors when scanning for QTL. CIM was performed in 10 cm windows using 10 markers as co-factors and significantly narrowed the windows around significant QTL peaks identified with SIM. BIM was used to identify the QTL model with the highest posterior density among models estimated in a Monte Carlo Markhov

Chain. BIM was performed using 100K steps in the MCMC chain with a burn-in of 20K steps. BIM is able to estimate the most likely number of QTL, their positions, and epistatic interactions between them. However, it cannot perform these estimates accurately when the X chromosome is included in males. Therefore, only the autosomes (chromosomes 2 and 3) were included in the BIM analysis.

The significance of the QTL models found in using SIM, CIM, and BIM was assessed using multiple interval mapping, which drops one QTL at a time to calculate the probability of QTL locations and their interactions using an ANOVA (Sen and Churchill, 2001). I used effect plots to understand direction of epistemic interactions between QTL.

Results

The sequencing effort resulted in 19.8-33.4 million reads in backcross libraries and to 41.7-48.5 million reads in parental libraries (Table 46). After aligning parental reads, calling variants in parental individuals (requiring min. 5x coverage), and merging variants among all members of a parental species (requiring 5% of individuals to share a variant), 295,982 *An. coluzzii* variants were used to subset the merged *An. coluzzii* "super parent" alignment file. Similarly, 477,959 variants were used to subset the merged *An. quadriannulatus* "super parent" alignment file. STACKS called fixed SNPs in these parental files and found 242,453 loci that were diagnostic between the species. STACKS analyzed 839 CQxQ males. Two males did not have good enough sequencing quality to be analyzed. After imposing a minimum threshold of 10 individuals per marker, stacks wrote genotypes for 821 individuals at 81,007 loci. While converting this dataset to R/qtl format in R, I removed loci that did not map to chromosomes 2, 3, or X (mtDNA, Y-linked, and unmapped loci). Thus, the initial input into R/qtl was 79,576 markers for 821 individuals. After filtering this dataset for markers genotyped in at least 120 individuals, and subsequently requiring individuals to be genotyped at a minimum of 10 markers, the dataset contained genotypes for 529 individuals at 796 markers. The loss of markers and individuals due to the fact that while an individual may have had

many markers for which genotypes were called, not all markers that were genotyped were shared among individuals.

After dropping additional markers and removing genotypes with high probability errors, I used the dropSimilarmarkers command of R/qtlTools to remove redundant markers on the autosomes. The final dataset from which the genetic map was estimated, and QTL analysis was performed was comprised of 421 CQxQ males genotyped at a 26% rate (prior to imputation) across 147 markers on chromosome 2, 113 on chromosome 3, and 66 on the X chromosome (Figure 26). The X chromosome markers are almost completely redundant; interestingly, one male has a recombinant X chromosome. The phenotypes of the individuals in this analysis (prior to culling individuals that were not well genotyped and in the final set) are highly skewed toward a sterility score of one, which is expected (Figure 27). All markers in this analysis show significant levels of segregation distortion (expected allele frequency of 0.5, chi-square p-value < 0.05, Figure 27). This is expected considering that inviability can occur in hybrids between member species of this complex (Slotman *et al*., 2015), and this pattern has been observed previously in backcrosses between species in the *An. gambiae* species complex (Slotman *et al.,* 2004, Smith *et al.*, 2015).

The second chromosome of the genetic map used in the QTL analysis is 159.9 centimorgans (cM) in length. The third is 152.9 cM, and the X is 0.4 cM (Figure 27). SIM using the Haley-Knott regression is a one-dimensional QTL scan that is only capable of identifying one QTL per linkage group (chromosome). If multiple QTL of large effect are located on one chromosome, the LOD ratio at markers between the two QTL can be inflated. The permutation test of this estimate calculated a 5% confidence threshold of QTL peaks that exceed a LOD score of 2.16, and a 1% confidence threshold of QTL peaks that exceed a LOD score of 2.93. Figure 28 shows the distribution of LOD scores for each marker position on the genetic map. Not surprisingly, the X chromosome is highly significant (upper right corner of the graph Figure 28). The results of SIM identified two wide QTL peaks on the second and third chromosomes that have significant LOD scores at the 5% and 1% confidence thresholds. These peaks can be

better visualized in plots of the second and third chromosomes in Figure 29. Solid red lines indicate the distribution of LOD scores across the genetic map. Horizontal dotted red lines indicate the 5% confidence threshold, and vertical dotted red lines indicate the location of the marker with the highest LOD score associated with these peaks. The 95% confidence estimates of QTL location on both chromosome 2 and chromosome 3 encompass over 50% of each chromosome. So, while it is evident there is at least one QTL with a significant effect on each chromosome, their locations are not precise. Despite this, I used multiple imputation to test the significance of this QTL model using the marker locations identified (X at 0 cM, chromosome 2 at 108 cM, and chromosome 3 at 101.4 cM). This model was significant (p-value < 0.05), had a LOD score of 89.69, and accounted for 62.49% of the variance observed in the phenotype. Of this, the X chromosome accounted for 53.84% of the observed variance. This analysis found significant effects of the X and 2nd chromosome QTL alone, significant interactions between all marker pairs, and the trio of all three markers. However, it did not find a sign cannot effect of the chromosome 3 QTL alone.

The results of the composite interval mapping identified QTL on the second, third, and X chromosomes, but only the X chromosome QTL was significant at the 5% confidence level identified through 10K permutations of this estimate. The lines plotted in blue on Figure 29 represent the results of the genome-wide scan of LOD scores (solid blue), the marker associated with the highest LOD score (vertical dotted blue), and the 5% confidence threshold. Despite the absence of significant QTL on the autosomes in this analysis, the peak on the third chromosome seems to give a more refined estimate of the chromosome 3 QTL location (Figure 29). I used multiple imputation to test the significance of this QTL model using the marker locations identified through CIM (X at 0 cM, chromosome 2 at 156.1 cM, and chromosome 3 at 98.0 cM). This model was also significant (p-value < 0.05), but had a slightly lower LOD score (88.66), and accounted for slightly less of the variance observed in the phenotype (62.08%) than the SIM model (Table 48).

The results of the Bayesian Imputation mapping found the model with the highest posterior density was one that involved five QTL across the autosomes (Figure 30). No prior was supplied to this simulation. The highest posterior density indicates that the MCMC chain spent the highest number of iterations in this model space. The five QTL model is supported by Bayes factor ratios calculated for models with numbers of QTL ranging from 1 to 13. The bottom panel of Figure 30 indicates that when the MCMC is in model space below five QTL the posterior probability of the model exceeds the prior (models with fewer QTL). As the MCMC chain moves past models with five QTL, the prior exceeds the posterior, and the MCMC chain returns to the five QTL model space. BIM cannot accurately incorporate the male hemizygous X chromosome into its models, so the results provided by this method need to be interpreted in a way that considers additional epistatic interactions between the autosomes and the X. The effects of the five autosomal QTL include both main and epistatic interactions between loci (Figure 31). The loci with the largest Bayes factor support are those which were previously identified by CIM on chromosomes 2 and 3. The QTL on chromosome 3 has a large main effect, and minor epistatic interactions with chromosome 2 loci. It is clear that the large QTL on chromosome 2 identified by SIM is actually comprised of four QTL. Chromosome 2 loci differ in their main and epistatic interactions. The QTL furthest to the left on the chromosome 2 graph (it is located on 2R) has almost all of its effect caused by epistasis, likely with the large locus on 2L (the furthest to the right). Loci whose sum effect (black lines) are not entirely explained by the main and epistatic effects graphed are likely epistatic with the X chromosome.

I used multiple imputation to test the significance of this QTL model using the marker locations identified at the highest LOD peaks. The full model is significant, (p-value < 0.05), has the highest LOD score observed at 108.69, and accounts for 69.5% of the variance in the phenotype (Table 49). Interestingly, however, significant effects are only observed for the X chromosome and the third chromosome QTL, and no interactions between QTL are significant. When the two minor effect QTL on the second chromosome are removed from this model (low LOD scores in Figure 31), the model is

still significant, the LOD score falls to 96.9% and the amount of variance explained drops to 65.6%. However, significant results are found at all QTL, and all pair-wise comparisons between two, three, and four QTL.

In order to better understand how these four loci (2R, 2L (largest LOD score), 3L, and X) are interacting with each other I compared the effect of their genotypes on sterility in a combined effect plot (Figure 32). This figure illustrates the huge effect inheritance of the *An. coluzzii* X chromosome has on sterility. When the *An. quadriannulatus* X is inherited the highest mean sterility observed is when both 2L and 3L are heterozygous. These two have a positive, epistatic effect on the phenotype; sterility increases to a level higher than the sum effect of these two loci if they are heterozygotes alone. When each autosomal QTL is heterozygous alone, it has only a minor effect on the phenotype. 2R has a negative epistatic interaction with 3L. This is particularly evident when the X chromosome is inherited from *An. coluzzii.* When this is the case, sterility is highest when wither 3L or 2R are the only heterozygous QTL. However, when both are heterozygous sterility decreases. This is not the case when a heterozygous 2R interacts with 2L, epitasis is positive.

I analyzed the locations of the identified QTL peaks in relation to significant QTL peaks found in the (*An. coluzzi* x *An. arabiensis*) x *An. arabiensis* cross (Slotman *et al.*, 2004) to see if the same regions of the *An. coluzzii* genome cause sterility when introgressed into *An. arabiensis* and *An. quadriannulatus* backgrounds. Two of the four significant QTL identified by Slotman *et al*. (2004) fall within the QTL peaks found in this study (Figure 33). These correspond to the 2L and 3L peaks I have discussed that were identified in the CIM and BIM analyses. I explored what genes could be responsible for the sterility phenotype in these regions by sub-setting the list of sex-biased genes in QUAD by those that fall under these QTL. I found 23, sex-biased, genes within this region that are significantly mis-expressed in COLZ x QUAD hybrid males in comparison to both QUAD and COLZ parental strains (Figure 34, Table 51). Interestingly, 16/23 genes show over-expression in comparison to the parental strains, whereas seven genes show under-expression in comparison to one or both parental

strains (Figure 34). Four of these genes have functions that are obviously related to reproduction. AGAP004221 is a CUP protein involved in female germ-line development and meiosis in *Drosophila*. AGAP005756 (cricklet) is involved in male mating behavior in *Drosophila*. AGAP006432 and AGAP006795 are peritrophins that have chitin binding properties and "major sperm protein" domains. Lastly, AGAP011032 is involved in cillium movement and axoneme assembly (the primary component of the sperm flagellum) (Table 51).

Discussion

The genotyping protocol yielded a large number of markers per individual that ultimately could not be used because they either did not overlap with the parental catalog, or were not shared amongst other CQxQ hybrids. This observation could be due to lower sequencing coverage per-individual in the backcross mosquitoes, or simply because of the large read depth distribution in the "super parent" alignments. This approach created a very large catalog of loci, which could result in almost all loci identified in hybrids being genotyped. These were later trimmed from the dataset to only include markers for which a larger proportion of individuals had been genotyped.

As expected, the results of the composite interval mapping inflated the LOD scores of regions surrounding major QTLs. This was very prevalent on the second chromosome, where two QTLs of smaller effect that lay between two large effect QTL were masked due to the inflated LOD scores. The Bayesian interval mapping analysis identified five autosomal loci that vary in their epistatic interactions. These results are drastically different from Slotman *et al*. (2004) who identified four significant autosomal QTL that interact with the X chromosome to cause sterility in *An. coluzzi* x *An. arabiensis* male hybrids. Similar to this study, the X chromosome in the *An. coluzzi* x *An. quadriannulatus* cross plays a major role in the sterility phenotype. In all QTL models tested, the X chromosome accounted for ~75-80% of the variation observed in the sterility phenotype. Unfortunately, due to the *Xag* inversion, we are not able to map genes to this chromosome. In previous studies focused on identifying sterility QTL in

Drosophila, researchers were only able to identify autosomal QTL that explained less than 2% of the variance in the phenotype (Moehring *et al.*, 2006). By removing the X from this analysis during Bayesian Interval mapping I was able to isolate the effect of the autosomes on the phenotype. As autosomal loci were added to the model tested using multiple interval mapping, LOD scores increased, and a higher percentage of the variation in the phenotype was explained. Additionally, autosomal QTL identified with this method narrowed and allowed a better understanding of their genomic location and interactions.

The results of the Bayesian interval mapping analysis identified QTL 2R as having an almost entirely epistatic effect (Figure 31). These results of the multiple interval mapping analysis in Table 50 indicate that in this reduced, but significant, model, all autosomal loci show significant epistatic interactions with each other and the X chromosome. The effect plots indicate that 2R has a negative epistatic interaction with 3L, where the impact of a heterozygous 3L QTL on the sterility phenotype is reduced by a heterozygous 2R QTL.

Due to the prevalence of epistatic interactions between QTL, it is important to understand how each effects the phenotype independently. By comparing the effects of each QTL on phenotype alone and in the presence of other heterozygous QTL (Figure 32), I found that autosomal loci have a minimal affect on the phenotype when acting alone. In contrast the interactions between QTLs 2L and 3L can significantly effect sterility in the absence of the *An. coluzzii* X. It is important to further explore the relationship between the overlapping *An. coluzzi* x *An. quadriannulatus* and *An. coluzzii* x *An. arabiensis* QTLs. These similarities in genomic regions that cause sterility in these crosses may be helpful in identifying genes that are rapidly evolving among species in the complex and integral to their ongoing divergence.

The mis-expressed COLZ x QUAD male genes found in the QTL regions had a higher proportion over-expression in comparison to parental males. This is not a common theme among mis-expressed genes between F1 hybrids and their parental strains (Chapter 3). If this is a common theme among genes that are implicated in

sterility it could help us narrow down causative genes in other crosses. Two genes involved in reproduction were found to be under-expressed in comparison to parental males: a gene involved in the development of sperm flagellum axoneme, and a CUP protein, which is involved in *Drosophila* germ-cell development and oogenesis. It is interesting that this gene is under-expressed in hybrid males in comparison to male parents. While all sex-biased genes (male- and female-) were included when analyzing mis-expression in the QTL windows, this gene is male biased between parental QUAD males and females. It would be interesting to explore why a gene that is involved in regulating oogenesis is upregulated in QUAD males and mis-expressed in hybrids.

In this study I have narrowly defined autosomal QTL that, in addition to the X chromosome, contribute to a large portion of the hybrid sterility phenotype in a cross between *An. coluzzii* and *An. quadriannulatus*. The shared QTL regions between the CxA and CxQ crosses suggest that these are important autosomal regions involved in the divergence of *An. coluzziii* from its sister species. Future analyses into factors contributing female sterility in the *An. coluzzi* x *An. quadriannulatus* cross will add to our knowledge of speciation genetic in this complex, and married with the hybrid gene expression data, may be able to identify additional genes that contribute to the hybrid phenotype.

REFERENCES

Aboagye-Antwi F, Alhafez N, Weedall GD, Brothwood J, Kandola S, Paton D, Fofana A, Olohan L, Betancourth MP, Ekechukwu NE, Baeshen R, Traore SF, Diabate A, Tripet F (2015) Experimental Swap of *Anopheles gambiae*'s Assortative Mating Preferences Demonstrates Key Role of X-Chromosome Divergence Island in Incipient Sympatric Speciation. *PLoS Genetics*, **11**, e1005141.

Anders S, Huber W (2010) Differential expression analysis for sequence count data. *Genome Biology*, **11**: R106.

Athrey G, Hodges TK, Reddy MR, Overgaard HJ, Matias A, Ridl FC, Kleinschmidt I, Caccone A, Slotman MA (2012) The Effective Population Size of Malaria Mosquitoes: Large Impact of Vector Control. *PLoS Genetics*, **8**, e1003097.

Baker DA, Russel S (2011) Role of testis-specific gene expression in sex-chromosome evolution of Anopheles gambiae. *Genetics* **189**:1117-1120.

Bashaw GJ, Baker BS (1997) The regulation of the *Drosophila msl-2* gene reveals a function for *Sex-lethal* in translational control. *Cell* **89**: 789-798.

Bayes JJ, Malik HS (2009) Altered heterochromatin binding by a hybrid sterility protein in *Drosophila* sibling species. *Science* **326**: 1538-1541.

Benjamini Y, Hockberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**: 289-300.

Besansky NJ, Krzywinski J, Lehmann T, Simard F, Kern M, Mukabayire O, Fontenile D, Toure Y, Sagnon NF (2003) Semipermeable species boundaries between *Anopheles gambiae* and *Anopheles arabiensis*: Evidence from multilocus DNA sequence variation. *PNAS* **100**: 10818-10823.

Besansky NJ, Powell JR, Caccone A, Hamm DM, Scott JA, Collins FH (1994) Molecular phylogeny of the *Anopheles gambiae* complex suggests genetic introgression between principal malaria vectors. *Proc Natl Acad Sci USA* **91**: 6885–6888.

Blatch S, Meyer KW, Harrison JF (2010) Effects of dietary folic acid level and symbiotic folate production on fitness and development in the fruit fly *Drosophila melanogaster*. *Fly* **4**: 312-319.

Bogh C, Lindsay SW, Clarke SE, Dean A, Jawara M, Pinder M, Thomas CJ (2007) High spatial resolution mapping of malaria transmission risk in The Gambia, West Africa using TM satellite imagery. *The American Journal of Tropical Medicine and Hygiene*, **76**: 875-881.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics* **30**: 2114-2120.

Bryan JH (1983) *Anopheles gambiae* and *Anopheles melas* at Brefet, The Gambia, and their role in malaria transmission. *Annals of Tropical Medicine and Parasitology* **77**:1-2.

Bryan JH, Petrarca V, Di Deco MA, Coluzzi M (1987) Adult behavior of members of the *Anopheles gambiae* complex in the Gambiae with special reference to *An. melas* and its chromosomal variants. *Parassitologia* **29**: 221-249.

Busing FMTA, Meijer E, Van Der Leeden R (1999) Delete-m jackknife for unequal m. *Statistics and Computing* **9**: 3-8.

Canty A, Ripley B (2015) boot: Bootstrap R (S-plus) functions. R package version 1.3-1.7.

Caputo B, Nwakanma D, Jawara M, Adiamoh M, Dia I, Konate L, Petrarca V, Conway DJ, della Torre A (2008) *Anopheles gambiae* complex along The Gambia river, with particular reference to the molecualr forms of *An. gambiae* s.s. *Malaria Journal* **7**: 182.

Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology* **22**: 2124-3140.

Catron DJ, Noor MAF (2008) Gene expression disruptions of organism versus organ in Drosophila species hybrids. *PLoS One* **3**: e3009.

Cheng C, White BJ, Kamdem C, Mockaitis K, Constantini C, Hahn MW, *et al.* (2012) Ecological Genomics of *Anopheles gambiae* along a latitudinal cline: a population-resequencing approach. *Genetics* **190**: 1417-1432.

Clark ME, O'Hara FP, Chawla A, Werren JH (2010) Behavioral and spermatogentic hybrid male breakdown in *Nasonia*. *Heredity* **104**: 289-301.

Clarkson CS, Weetman D, Essandoh J, Yawson AE, Maslen G, Manske M, Field SG, Webster M, Antão T, MacInnis B, Kwiatkowski D, Donnelly MJ (2014) Adaptive introgression between *Anopheles* sibling species eliminates a major genomic island but not reproductive isolation. *Nature Communications* **5**: 4248.

Clements AN: Spermatogenesis and the structure of spermatozoa. The biology of mosquitoes, Development, Nutrition and Reproduction. Edited by: Clements AN. 1992, London: Chapmann & Hall, **1**: 333-335.

Cline TW (1993) The Drosophila sex determination signal: how do flies count to two? *Trends in Genetics* **9**: 385-390.

Coetsee M, Hunt RH, Wilkerson R, della Torre A, Coulibaly MB, Besansky NB (2013) *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa* **3619**: 246-274.

Coetzee M, Craig M, le Sueur D (2000) Distribution of African Malaria Mosquitoes Belonging to the *Anopheles gambiae* Complex. *Parasitology Today* **16**: 74-77.

Coetzee M, Hunt RH, Wikerson R, della Torre A, Coulibaly MB, Besansky NJ (2013) *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa* **3619**: 246-274.

Cohuet A, Krishnakumar S, Simard F, Morlais I, Koutsos A, Fontenille D, Mindrinos M, Kafatos FC (2008) SNP discovery and molecular evolution in *Anopheles gambiae*, with special emphasis on innate immune system. *BMC Genomics* **9**: 227.

Coluzzi M, Sabatini A, della Torre A, Di Deco MA, Petrarca V (2002) A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science* **298**: 1415–1418.

Coolon JD, and Wittkopp PJ (2013) cis- and trans-Regulation in *Drosophila* Interspecific Hybrids. In: Polyploid and Hybrid Genomics, pg. 37-57. Wiley & Sons, eds. Chen ZJ, Birchler JA.

Cowles C, Hirschhorn J, Altschuler D, Lander E (2002) Detection of regulatory variation in mouse genes. *Nature Genetics* **32**: 432-437.

Cutter AD (2008) Divergence times in *Caenorhabditis* and *Drosophila* inferred from direct estimates of the neutral mutation rate. *Mol Biol Evol* **25**: 778–786.

Darum SR (2006) Beyond Bonferroni: Less conservative analyses for conservation genetics. *Conserv Genet* **7**: 783-787.

Davidson G (1962) *Anopheles gambiae* complex. *Nature* **196**: 907.

Davison AC, Hinkley DV (1997) Bootstrap methods and their applications. Cambridge: Cambridge University Press.

Deitz KC, Athrey G, Reddy MR, Overgaard HJ, Matias A, Jawara M, *et al.* (2012) Genetic isolation within the malaria mosquito *Anopheles melas*. *Mol Ecol* **18**: 4498-4513.

Deitz KC, Athrey GA, Jawara M, Overgaard HJ, Matias A, Slotman MA (2016) Genome-Wide Divergence in the West-African Malaria Vector Anopheles melas. *G3: Genes, Genomes, Genetics* **8**: 2867-2879.

Della Torre A, Fanello C, Akogbeto M, Dossou-yovo J, Favia G, Petrarca V, *et al.* (2001) Molecular evidence of incipient speciation within *Anopheles gambiae* s.s. in West Africa. *Insect Mol Biol* **10**: 9-18.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39**: 1-38.

Diabaté A, Dabire RK, Millogo N, Lehmann T (2007) Evaluating the effect of postmating isolation between molecular forms of *Anopheles gambiae* (Diptera: Culicidae). *J Med Entomol* **44**: 60-64.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Dobzhansky T (1937) Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics* **21**: 113-135.

Donelly MJ, Townson H (2000) Evidence for extensive genetic differentiation among populations of the malaria vector *Anopheles arabiensis* in Eastern Africa. *Insect Mol Biol* **9**: 357-367.

Dunn OJ (2012) Multiple comparison using rank sums. *Technometrics* **6**: 241-252.

Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. *Mol Biol Evol* **28**: 2239-2252.

Erickson JW, Quintero JJ (2007) Indirect Effects of Ploidy Suggest X Chromosome Dose, Not the X:A Ratio, Signals Sex in *Drosophila*. *PLoS Biol* **5**: e332.

Fabian DK, Kapun M, Nolte V, Kofler R, Schmidt PS, Schlötterer C, Flatt T (2012) Genome-wide patterns of latitudinal differentiation among populations of *Drosophila melanogaster* from North America. *Molecular Ecology* **21**: 4748-4769.

Favia G, Lanfrancotti A, Spanos L, Siden-Kiamos I, Louis C (2001) Molecular characterization of ribosomal DNA polymorphisms discriminating among chromosomal forms of *Anopheles gambiae* s.s. *Insect Mol Biol* **10**: 19–23.

Ferree PM, Barbash DA (2009) Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in *Drosophila*. *PLoS Biology* **7**: e1000234.

Fisher S, Barry A, Abreu J, Minie J, Delorey TM, Young G, Fennell TJ, Allen A, Ambrogio L, Berlin AM, Blumenstiel B, Cibulskis K, Friedrich D, Johnson R, Juhn F, Reilly B, Shammas R, Stalker J, Sykes SM, Thompson J, Walsh J, Zimmer A, Zwirko Z, Gabriel S, Nicol R, Nasbaum C (2011) A scaleable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biology **12:** R1.

Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, *et al.* (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science* **347**: 1258524-1-6.

Gagon C (1995) Regulation of Sperm Motility at the Axonemal Level. *Reproductive Fertility and Development* **7**: 847-855.

Gao S, Giansanti MG, Buttrick GJ, Ramasubramanyan S, Auton A, Gatti M, Wakefield JG (2008) Australin: a chromosomal passenger protein required specifically for *Drosophila melanogaster* male meiosis. *Journal of Cell Biology* **180**: 521-535.

Garrigan D, Kingan SB, Geneva AJ, Andolfatto P, Clark AG, Thornton KR, *et al.* (2012) Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Res* **22**: 1499-1511.

Gelbart ME, Kuroda MI (2009) *Drosophila* dosage compensation: a complex voyage to the X chromosome. *Development* **136**: 1399-1410.

Gelfand HM (1955) *Anopheles gambiae* Giles and *An. melas* Theobald in a coastal area of Libera, West Africa. *Trans R Soc Trop Med Hyg* **49**: 508-527.

Gentile G, Slotman M, Ketmaier V, Powell JR, Caccone A (2001) Attempts to molecularly distinguish cryptic taxa in *Anopheles gambiae* s.s. *Insect Mol Biol* **10**: 25–32.

Gergen JP (1987) Dosage compensation in *Drosophila*: Evidence that *daughterless* and *Sex-lethal* control X chromosome activity at the blastoderm stage of embryogenesis. *Genetics* **117**: 477-485.

Giraldo-Calderon GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, Ho N, Gesing S, VectorBase Consortium, Madey G, Collins H, Lawson D (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Research* **43**: D707-D713.

Giraldo-Calderón GI, Emrich SJ, MacCallum RM, Maslen G, Dialynas E, Topalis P, Ho N, Gesing S, VectorBase Consortium, Madey G, Collins FH, Lawson D (2015) VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Res* **43** (Database issue): D707-D713.

Gomes S and Civetta A (2014) Misregulation of spermatogenesis genes in Drosophila hybrids is lineage-specific and driven by the combined effects of sterility and fast male regulatory divergence. *Journal of Evolutionary Biology* **27**: 1775-1783.

Gomes S, Civetta A (2015) Hybrid male sterility and genome-wide misexpression of male reproductive proteases. *Scientific Reports* **5**: 11976.

Gramates LS, Marygold SJ, dos Santos G, Urbano J-M, Antonazzo G, Matthews BB, Rey AJ, Tabone CJ, Crosby MA, Emmert DB, Falls K, Goodman JL, Hu Y, Ponting L, Schroeder AJ, Strelets VB, Thurmond J, Zhou P and the FlyBase Consortium (2017) FlyBase at 25: looking to the future. *Nucleic Acids Research* **45**: D663-D671.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, *et al.* (2010) A draft sequence of the neandertal genome. *Science* **328**: 710-722.

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315-324.

Hall AB, Basu S, Jiang X, Qi Y, Timoshevskiy VA, Biedler JK, Sharakhova MV, Elahi R, Anderson MAE, Chen XG, Sharakhov IV, Adelman ZN, Tu Z (2015) A male-determining factor in the mosquito *Aedes aegypti*. *Science* **348**: 1268-1270.

Hall AB, Papathanos PA, Sharma A, Cheng C, Akbari OZ, Assour L, Bergman NH, Cagnetti A, Crisanti A, Dottorini T, Fiorentini E, Galizi R, Hnath J, Jiang X, Koren S, Nolan T, Radune D, Sharakhova MV, Steele A, Timoschevskiy VA, Windbichler N, Zhang S, Hahn MW, Phillippy AM, Emrich SJ, Sharakhov IV, Tu ZJ, Besansky NJ (2016) Radical remodeling of the Y chromosome in a recent radiation of malaria mosquitoes. *PNAS* **113**: E2114-E2123.

Hilfiker A, Amrein H, Dubendorfer A, Schneiter R, Nothiger R (1995) The gene virilizer is required for female-specific splicing controlled by Sxl, the master gene for sexual development in *Drosophila*. *Development* **121**: 4017-4026.

Holt R, Subramanian G, Halpern A, Sutton G, Charlab R, *et al.* (2002) The genome sequence of the malaria mosquito *Anopheles gambiae*. *Science* **298**, 129-149.

Hunt RH, Coetzee M, Fettene M (1998) The *Anopheles gambiae* complex: a new species from Ethiopia. *Trans Royal Soc Trop Med & Hygiene* **92**: 231-235.

Jiang H, Wong WH (2009) Statistical inferences for isoform expression in RNA-seq. *Bioinformatics* **25**: 1026-1032.

Jiang X, Biedler JK, Qi Y, Hall AB, Tu Z (2015) Complete dosage compensation in *Anopheles stephensi* and evolution of sex-biased genes in mosquitoes. *Genome Biology and Evolution* **7**: 1914-1924.

Karlsen BO, Klingan K, Emblem A, Jorgensen TE, Jueterbock AJ, Furmanek T, *et al.* (2013) Genomic divergence between migratory and stationary ecotypes of Atlantic cod. *Mol Ecol* **22**: 5098-5011.

Kharchenko PV, Xi R, Park PJ (2011) Evidence for dosage compensation between the X chromosome and autosomes in mammals. *Nature Genetics* **43**: 1167-1169.

Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, *et al.* (2011a) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. *PLoS One* **6**: e15925.

Kofler R, Pandey RV, Schlotterer C (2011b) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). *Bioinformatics* **27**: 3435-3436.

Korneliussen TS, Albrechtsen A, Nielsen R (2014) ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**: 356.

Kosambi DD (1943) The estimation of map distances from recombination values. *Annals of Eugenics* **12**: 172-175.

Kruskal WH, Wallis WA (2012) Use of ranks in on-criterion variance analysis. *Journal of the American Statistical Association* **47**: 583-621.

Krzywinska E, Dennison NJ, Lycett GJ, Krzywinski J (2016) A maleness gene in the malaria mosquito *Anopheles gambiae*. *Science* **353**: 67-69.

Krzywinska E, Krzywinski J (2009) Analysis of expression in the *Anopheles gambiae* developing testes reveals rapidly evolving lineage-specific genes in mosquitoes. *BMC Genomics* **10**: 300.

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185-199.

Landry, CR, Wittkopp, PJ, Taubes, CH, Ranz, JM, Clark, AG, and Hartl, DL (2005) Compensatory cis-trans evolution and the dysregulation of gene expression in interspecific hybrids of Drosophila. *Genetics* **171**: 1813-1822.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

Lanzaro GC, Lee Y (2013) Speciation in *Anopheles gambiae* - The Distribution of genetic polymorphism and patterns of reproductive isolation among natural populations. In: Manguin S (ed) *Anopheles mosquitoes - New insights in to malaria vectors*.

Lee Y, Marsden CD, Norris LC, Collier TC, Main BJ, Fofana A, Cornel AJ, Lanzaro GC (2013) Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, *Anopheles gambiae*. *PNAS* **110**: 19854-19859.

Lehman T, Licht M, Elissa N, Maega BT, Chimumbwa JM, Watsenga FT, *et al.* (2003) Population structure of *Anopheles gambiae* in Africa. *J Hered* **94**: 133–147.

Lemos, B, Araripe, LO, Fontanillas, P, and Hartl, DL (2008) Dominance and the evolutionary accumulation of cis- and trans- effects on gene expression. *PNAS* **105**: 14471-14476.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.* (2009) The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N., Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**: 2078-9.

Lim C, Gandhi S, Biniossek ML, Feng L, Schilling O, Urban S, Chen X (2015) An Aminopeptidase in the Drosophila Testicular Niche Acts in Germline Stem Cell Maintenance and Spermatogonial Dedifferentiation. *Cell Reports* **13**: 315-324.

Linck, RQ, Chemes, H, Albertini D (2016) The axoneme: the propultive engine of spermatozoa and cilia and associated ciliopathies leading to infertility. *Journal of Assisted Reproduction and Genetics* **33**: 141-156.

Loaiza JR, Bermingham E, Sanjur OI, Scott ME, Bickersmith SA, Conn JE (2012) Review of genetic diversity in malaria vectors (Culicidae: Anophilinae). *Infect Genet Evol* **12**: 1-12.

Lucchesi JC, Kelly WG, Panning B (2005) Chromatin remodeling in dosage compensation. *Annual Reviews Genetics* **39**: 615-651.

Lucchesi JC, Skripsky T (1981) The link between dosage compensation and sex differentiation in *Drosophila melanogaster*. *Chromosoma* **82**: 217-227.

Lunther G, Goodson M (2011) Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research* **21**: 936-939.

Magnusson K, Lycette GJ, Mendes AM, Lynd A, Papathanos PA, Crisanti A, Windbichler N (2012) Demasculization of the *Anopheles gambiae* X chromosome. *BMC Evol Biol*, **12**: 69.

Maheshwari S, Barbash DA (2012) An indel polymorphism in the hybrid incompatibility gene *Lethal Hybrid Rescue* of *Drosphila* is functionally relevant. *Genetics* **192**: 683-691.

Maheshwari S, Barbash DA (2012) Cis-by-trans regulatory divergence causes the asymmetric lethal effects of an ancestral hybrid incompatibility gene. *PLoS Genetics* **8**: e1002597.

Mallet J (1995) A species definition for the Modern Synthesis. *TREE* **10**: 294-299.

Mallet J, Besansky N, Hahn MW (2015) How reticulated are species? *BioEssays* **38**: 140-149.

Marchand RP (1983) Field observations on swarming and mating in *Anopheles gambiae* mosquitoes in Tanzania. *Neth J Zool* **34**: 367-387.

Marsden CD, Lee Y, Kreppel K, Weakley A, Cornel A, Ferguson HM, Eskin E, Lanzaro GC (2014) Diversity, Differentiation, and Likage Disequilibrium: Prospects for Association Mapping in the Malaria Vector *Anopheles arabiensis*. *G3: Genes, Genomes, Genetics* **4**: 121-131.

Marsden CD, Lee Y, Nieman CC, Sandford MR, Dinis J, Martins C, *et al.* (2011) Asymmetric introgression between the M and S forms of the malaria vector,

Anopheles gambiae, maintains divergence despite extensive hybridization. *Mol Ecol* **20**: 4983–4994.

Mayr E (1970) Populations, Species, and Evolution. Belknap Press of Harvard University Press: Cambridge, MA and London, England.

McManus CJ, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, Wittkopp J (2010) Regulatory divergence in *Drosophila* revealed by mRNA-seq. *Genome Research* **20**: 816-825.

McManus, CJ, Coolon, JD, Duff, MO, Eipper-Mains, J, Graveley, BR, and Wittkopp, PJ (2010) Regulatory divergence in Drosophila revealed by mRNA-seq. *Genome Research* **2:** 816-825.

Meiklejohn CD, Presgraves DC (2012) Little evidence for demasculinization of the Drosophila X chromosome among genes expressed in the male germline. *Genome Biol Evol* **4**: 1007-1016.

Mi H, Dong Q, Muruganujan A, Gaudet P, Lewis S, Thomas PD (2010) PANTHER version 7: improved phylogenetic trees, orthologs, and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res* **38**: D204-D210.

Michalak P, Noor MAF (2003) Genome-wide patterns of expression in *Drosophila* pure species and hybrid males. *Molecular Biology and Evolution* **20**: 1070-1076.

Michalak P, Noor MAF (2004). Association of misexpression with sterility in hybrids of Drosophila simulans and D. mauritiana. *Journal of Molecular Evolution* **59**, 277-282.

Moehring AJ, Llopart A, Elwyn S, Coyne JA, Mackay TFC (2006) The genetic basis of postzygotic reproductive isolation between *Drosophila santomea* and *D. yakuba* due to hybrid male sterility. *Genetics* **173**: 225–233.

Moehring AJ, Teeter KC, Noor MAF (2007) Genome-wide patterns of expression in Drosophila pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. *Molecular Biology and Evolution* **24**: 137-145.

Montague MJ, Li G, Gandolfi B, Khan R, Aken BL, Searle SM, *et al.* (2014) Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. *Proc Natl Acad Sci U S A* **111**: 17230-17235.

Moreno M, Salgueiro P, Vicente JL, Cano J, Berzosa PJ, de Lucio A, Simard F, Caccone A, Do Rosario VE, Pinto J, Benito A (2007) Genetic population structure of *Anopheles gambiae* in Equatorial Guinea. *Malaria Journal* **6**: 137.

Nariai N, Kojima K, Mimori T, Kawai Y, Nagasaki M (2015) A Bayesian approach for estimating allele-specific expression from RNA-Seq data with diploid genomes. *BMC Genomics* **17**: 2.

Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, *et al.* (2015) Highly evolvable malaria vectors: the genomes of 16 *Anopheles* mosquitoes. *Science* **347**: 1258522.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction enzyme endonucleases. *Proc Natl Acad Sci U S A* **10**: 5269-5273.

Newton ME, Southern DI, Wood RJ (1974) X and Y chromosomes of *Aedes aegypti* (L.) distinguished by Giemsa C-banding. *Chromosoma* **49**: 41-19.

Norris LC, Main BJ, Lee Y, Collier TC, Fofana A, Cornel AJ, Lanzaro GC (2015) Adaptive introgression in an African malaria mosquito coincident with the increased usage of insecticide-treated bed nets. *Proc Natl Acad Sci USA* **112**: 815-820.

Nosil P (2012) Ecological Speciation. Oxford University Press: Oxford, England.

Nosil P, Schluter D (2011) The genes underlying the process of speciation. *Trends Ecol Evol* **4**: 160-167.

Okereke TA (1980) Hybridization studies on sibling species of the *Anopheles gambiae* Giles complex (Diptera, Culicidae) in the laboratory. *Bull Entomol Res* **70**: 391-398.

Orchard S, Ammari M, Aranda B, Breuza L, Briganti L, *et al.* (2014) The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Research* **42**: D358-D363.

Orr HA, Presgraves DC (2000) Speciation by postzygotic isolation: forces, genes and molecules. *BioEssays* **22**: 1085–1094.

Overgaard HJ, Reddy VP, Abaga S, Matias A, Reddy MR, Kulkarni V, Schwabe C, Segura L, Kleinschmidt I, Slotman MA (2012) Malaria transmission after five years of vector control on Bioko Island, Equatorial Guinea. *Parasites & Vectors*, **5**, 253.

Powell JR, Petrarca V, della Torre A, Caccone A, Coluzzi M (1999) Population structure, speciation, and introgression in the *Anopheles gambiae* complex. *Parassitologia* **41**: 101-113.

Powell THQ, Hood GR, Murphy MA, Heilveil JS, Berlocher SH, Nosil P, *et al.* (2013) Genetic divergence along the speciation continuum: the transition from host race to species in *Rhagoletis* (Diptera: Tephritidae). *Evolution* **67**: 2561-2576.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Redd MR, Overgaard HJ, Abaga S, Reddy VP, Caccone A, Kiszewski AE, Slotman MA (2011) Outdoor host seeking behavior of *Anopheles gambiae* mosquitoes following initiation of malaria vector control on Bioko Island, Equatorial Guinea. *Malaria Journal* **10:** 184.

Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J (2016) g:Profiler - a web server for functional interpretation of gene lists. *Nucleic Acids Research* **44**: W83-W89.

Rose G, Krzywinska E, Kim J, Revuelta L, Ferretti L, Krzywinski J (2016) Dosage compensation in the African malaria mosquito *Anopheles gambiae*. *Genome Biology and Evolution* **8**: 411-425.

Satyaki PRV, Cuykendall TN, Wei KH-C, Brideau NJ, Kwak H, Aruna S, et al. (2014) The *Hmr* and *Lhr* Hybrid Incompatibility Genes Suppress a Broad Range of Heterochromatic Repeats. *PLoS Genetics* **10**: e1004240.

Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals - mining genome-wide polymorphism data without big funding. *Nat Rev Genet* **15**: 749-763.

Sen S, Churchill GA (2001) A statistical framework for quantitative trait mapping. *Genetics* **159**: 371-387.

Sharakhov IV, White BJ, Sharakhova MV, Kayondo J, Lobo NF, Santolamazza F, della Torre A, Simard F, Collins FH, Besansky NJ (2006) Breakpoint structure reveals the unique origin of an interspecific chromosomal inversion (*2La*) in the *Anopheles gambiae* complex. *PNAS* **103**: 6258-6262.

Sharakhova MV, George P, Brunsentsova IV, Leman SC, Bailey JA, Smith CD, Sharakhov IV (2010) Genome mapping and characterization of the *Anopheles gambiae* heterochromatin. *BMC Genomics* **11**: 459.

Sharakhova MV, Hammond MP, Lobo NF, Krzywinski J, Unger MF, Hillenmeyer ME, *et al.* (2007) Update of the *Anopheles gambiae* PEST genome assembly. *Genome Biol* **8**: R5.

Simard F, Ayala D, Kamdem GC, Pombi M, Etouna J, Ose K, *et al.* (2009) Ecological niche partitioning between *Anopheles gambiae* molecular forms in Cameroon: the ecological side of speciation. *BMC Ecol* **9**: 17.

Singh R S, Kulathinal RJ (2000) Sex gene pool evolution and speciation: a new paradigm. *Genes Genet Syst* **75**: 119– 130.

Slotman MA, della Torre A, Calzetta M, Powell JR (2005a) Differential introgression of chromosomal regions between *Anopheles gambiae* and *An. arabiensis*. *The American Journal of Tropical Medicine and Hygiene* **73**: 326-335.

Slotman MA, della Torre A, Powell JR (2004) The genetics of inviability and male sterility in hybrids between *Anopheles gambiae* and *Anopheles arabiensis. Genetics* **167**: 275-287.

Slotman MA, della Torre A, Powell JR (2005b) Female sterility in hybrids between *Anopheles gambiae* and *An. arabiensis* and the causes of Haldane's rule. *Evolution* **59**: 1016-1026.

Smith HA, White BJ, Kundert P, Cheng C, Romero-Severson J, Andolfatto P, Besansky NJ (2015) Genome-wide QTL mapping of saltwater tolerance in sibling species of *Anopheles* (malaria vector) mosquitoes. *Heredity* **115**: 471-479.

Struchiner CJ, Slotman MA, Fontaine MC, Deitz KC, Love RR, Witzig C, *et al.* (2015) Multi-species comparison of genome-wide evolutionary signatures of distinct gene functional classes and chromosome among malaria vectors. *Insect Mol Biol*, in review.

Sundararajan V, Civetta A (2011) Male sex interspecies divergence and down regulation of expression of spermatogenesis genes in *Drosophila* sterile hybrids. *Journal of Molecular Evolution* **72**: 80-89.

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585-595.

Takahasi KR, Matsuo T, Takano-Shimizu-Kouno T (2011) Two types of cis-trans compensation in the evolution of transcription regulation. *PNAS* **108**: 15276-15281.

Temu EA, Hunt RH, Coetzee M, Minjas JM, Shiff CJ (1997) Detection of hybrids in natural populations of the Anopheles gambiae complex by the rDNA-based, PCR method. *Ann Trop Med Parasitol* **91**: 963–965.

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, *et al.* (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* **13**: 2129-2141.

Toure YT, Petrarca V, Traore SF, Coulibaly A, Maiga HM, Sankare O, Sow M, Di Deco MA, Coluzzi M (1998) The distribution and inversion polymorphism of chormosomally recognized taxa of the *Anopheles gambiae* species complex in Mali, West Africa. *Parassitologia* **40**: 477-511.

Tripet R, Thiemann T, Lanzaro GC (2005) Effect of seminal fluids in mating between M and S forms of *Anopheles gambiae*. *J Med Entomol* **42**: 596-603.

Tukey J (1949) Comparing Individual Means in the Analysis of Variance. *Biometrics* **5**: 99–114.

Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic Islands of Speciation in *Anopheles gambiae*. *PLoS Biol* **9**: e285.

Ulmscheider B, Grillo-Hill BK, Benitex M, Azimova DR, Barber DL, Nystul TG (2016) Increased intracellular pH is necessary for adult epithelial and embryonic stem cell differentiation. *Journal of Cell Biology* **215**: 345-355.

Vicoso B, Bachtrog D (2011) Lack of global dosage compensation in *Schistostoma mansoni*, a female-heterogametic parasite. *Genome Biol Evol* **3:** 230-235.

Weetman D, Steen K, Rippon EJ, Mawejje HD, Donelly MJ, Wilding CS (2014) Contemporary gene flow between wild *An. gambiae* s.s. and *An. arabiensis* (2014) *Parasit Vectors* **7**: 345.

White BJ, Cheng C, Simard F, Simard F, Constantini C, Besansky NJ (2010) Genetic association of physically unlinked islands of genomic divergence in incipient species of *Anopheles gambiae*. *Molecular Ecology* **19**: 925-939.

White BJ, Collins FH, Besansky NJ (2011) Evolution of *Anopheles gambiae* in relation to humans and malaria. *Annu Rev Ecol Evol Syst* **42**: 111-132.

White BJ, Santolamazza F, Kamau L, Pombi M, Grushko O, Mouline K, Brengues C, Guelbeogo W, Coulibaly M, Kayondo JK, Sharakhov I, Simard F, Petrarca V, della Torre A, Besansky NJ (2007) Molecular karyotyping of the *2La* inversion in *Anopheles gambiae*. *Am J Trop Med Hyg* **76**: 334-339.

White GB (1974) *Anopheles gambiae* complex and disease transmission in Africa. *Trans R Soc Trop Med Hyg* **68**: 278-298.

White GB (1974) Anopheles gambiae complex disease transmission in Africa. *Transactions of the Royal Society of Tropical Medicine & Hygiene* **68**: 278-298.

White-Cooper H, Doggett K, Ellis RE (2009) The evolution of spermatogenesis. In: Sperm Biology: An Evolutionary Perspective (TR Birkhead, DJ Hosken & S Pitnick, eds), 151–183. Academic Press, Burlington, MA.

Wilding C, Weetman D, Steen K, Donnelly M (2009) High, clustered, nucleotide diversity in the genome of *Anopheles gambiae* revealed through pooled-template

sequencing: implications for high-throughput genotying protocols. *BMC Genomics* **10**: 320.

Wittkopp PJ, Carroll SB, Kopp A (2003) Evolution in black and white: genetic control of pigment patterns in *Drosophila*. *Trends in Genetics* **19**: 495-504.

Wittkopp PJ, Haerum BK, Clark AG (2004) Evolutionary changes in cis and trans gene regulation. *Nature* **430**: 85-88.

Wray GA, Hahn MW, Abouheif E, Pizer M, Rockman MV, Romano LA (2003) The evolution of transcription in eukaryotes. *Molecular Biology and Evolution* **20**: 1377-1419.

Wu C-I, Perez DE, Davis AW, Johnson NA, Cabot EL, Palopoli MF, Wu M-L (1992) Molecular genetic studies of postmating reproductive isolation in Drosophila. In: Takahata N, Clark AG (eds) Molecular Paleo-population Biology. Springer-Verlag, Berlin, 191–212.

Yandell BS, Mehta T, Banerjee S, Shriner D, Venkataraman R, Moon JY, Neely WW, Wu H, Smith R, Yi N (2007) R/qtlbim: QTL with Bayesian Interval Mapping in experimental crosses. *Bioinformatics* **23**: 641-643.

APPENDIX A

TABLES

Chapter I

Table 1 Estimates of mean nucleotide diversity (π) and Tajima's *D* for each chromosome arm and *An. melas* population, measured in 100 kb, non-overlapping sliding windows. Values in parentheses indicate the standard error of the mean for each statistic. Regions of heterochromatin in the *An. gambiae* genome were removed from summary statistics. Reprinted from Deitz *et al.* (2016).

	X		2R		2L	
Population	Mean π	Mean Tajima's *D*	Mean π	Mean Tajima's *D*	Mean π	Mean Tajima's *D*
West	0.0046 (0.00008)	-0.100 (0.0050)	0.0045 (0.00009)	-0.126 (0.0029)	0.0053 (0.00009)	-0.108 (0.0032)
South	0.0035 (0.00010)	-0.035 (0.0054)	0.0045 (0.00010)	-0.030 (0.0024)	0.0050 (0.00010)	-0.025 (0.0026)
Bioko	0.0029 (0.00008)	-0.042 (0.0070)	0.0029 (0.00009)	-0.038 (0.0037)	0.0035 (0.00013)	-0.024 (0.0037)
	3R		3L		Genome-wide	
Population	Mean π	Mean Tajima's *D*	Mean π	Mean Tajima's *D*	Mean π	Mean Tajima's *D*
West	0.0058 (0.00013)	-0.107 (0.0028)	0.0058 (0.00011)	-0.093 (0.0028)	0.0050 (4.78×10^{-5})	-0.1092 (0.0014v
South	0.0053 (0.00014)	-0.032 (0.0025)	0.0051 (0.00012)	-0.026 (0.0025)	0.0048 (5.31×10^{-5})	-0.0291 (0.0012)
Bioko	0.0037 (0.00012)	-0.022 (0.0039)	0.0039 (0.00012)	-0.021 (0.0038)	0.0034 (5.12×10^{-5})	-0.0287 (0.0018)

Table 2 Summary statistics of the F_{ST} null distribution and false discovery rate simulation. Reprinted from Deitz *et al.* (2016).

	Population Pool Distribution	Sequencing Pool Distribution	Allele Frequency Difference Distribution	Pair-wise F_{ST} Distribution
Minimum	0.100	0.000	0.000	0.000
Q1	0.450	0.433	0.067	0.005
Median	0.500	0.500	0.100	0.020
Mean	0.500	0.500	0.135	0.046
Q3	0.550	0.567	0.200	0.060
Maximum	0.900	1.000	0.700	0.875

Table 3 Number of significant (Sig.) and fixed SNPs per chromosome in each pair-wise *An. melas* population comparison. Regions of heterochromatin in the *An. gambiae* genome were removed from summary statistics. Reprinted from Deitz *et al.* (2016).

	X		2R		2L		3R		3L		Genome-wide	
Comparison	Fixed	Sig.	Fixed	Sig.	Fixed	Sig.	Fixed	Sig.	Fixed	Sig.	Fixed	Sig.
West - South	879	3,028	185	3,853	202	3,624	116	3,340	220	3,272	1,602	17,117
West - Bioko	319	1,810	439	6,373	403	5,061	299	4,671	264	3,512	1,724	21,427
South - Bioko	1,725	4,324	981	10,396	1,110	9,197	692	8,825	879	6,988	5,387	39,730

Table 4 Results of sequence read trimming, mapping, and filtering. Raw sequence reads were trimmed using Trimmomatic ver. 0.35. Paired-end reads were trimmed to a minimum Phred quality score of 20, and a minimum length of 50 bp. Only reads with both surviving pairs were retained for mapping. Reads were mapped and filtered to exclude those with mapping quality values (MAPQ) less than 20. Reprinted from Deitz *et al.* (2016).

	Population	**West**				
	SRA Run Accession	SRR567657				
	Number Read Pairs	78,025,712				
	Read Pairs Post Trimming	71,441,506				
	Chromosome	X	2R	2L	3R	3L
	Reference Size (bp)	24,393,108	61,545,105	49,364,325	53,200,684	41,963,435
Raw Reads	Reads Mapped	14,738,106	25,737,921	20,480,713	20,907,372	16,558,122
	Mean Read Length (bp)	98.79	98.88	98.94	99.00	99.04
	Read Length SD (bp)	7.09	6.92	6.83	6.73	6.64
	Mean Reads / bp	57.69	40.54	40.17	38.09	38.17
	Reads/bp SD	1,600.83	250.59	102.95	60.78	573.04
Filtered Reads	Reads Mapped	9,270,784	22,656,116	17,723,120	18,286,869	13,696,759
	Mean Read Length (bp)	98.70	98.88	98.92	99.00	99.01
	Read Length SD (bp)	7.17	6.90	6.83	6.70	6.67
	Mean Reads / bp	36.56	35.76	34.85	33.39	31.65
	Reads/bp SD	814.74	177.19	28.70	28.11	160.39

	Population	**South**				
	SRA Run Accession	SRR567658				
	Number Read Pairs	52,594,743				
	Read Pairs Post Trimming	38,797,029				
	Chromosome	X	2R	2L	3R	3L
	Reference Size (bp)	24,393,108	61,545,105	49,364,325	53,200,684	41,963,435
Raw	Reads Mapped	8,063,847	12,799,679	10,236,490	10,365,790	8,216,090

Reads	Mean Read Length (bp)	98.89	98.96	99.03	99.08	99.11
	Read Length SD (bp)	6.92	6.78	6.65	6.58	6.50
	Mean Reads / bp	31.38	20.04	19.94	18.77	18.84
	Reads/bp SD	977.96	178.37	156.58	32.51	319.12
Filtered Reads	Reads Mapped	4,995,277	11,231,431	8,746,514	9,039,990	6,802,292
	Mean Read Length (bp)	98.84	98.96	99.01	99.08	99.10
	Read Length SD (bp)	6.91	6.74	6.66	6.55	6.50
	Mean Reads / bp	19.58	17.62	17.10	16.41	15.63
	Reads/bp SD	593.91	133.18	16.44	15.62	123.40

	Population	**Bioko**				
	SRA Run Accession	SRR606147				
	Number Read Pairs	56,776,632				
	Read Pairs Post Trimming	50,934,546				
	Chromosome	X	2R	2L	3R	3L
	Reference Size (bp)	24,393,108	61,545,105	49,364,325	53,200,684	41,963,435
Raw Reads	Reads Mapped	11,841,657	18,933,558	15,112,849	15,353,319	12,229,559
	Mean Read Length (bp)	98.80	98.87	98.93	98.99	99.03
	Read Length SD (bp)	7.05	6.92	6.83	6.73	6.64
	Mean Reads / bp	46.32	29.88	29.69	28.03	28.25
	Reads/bp SD	1534.00	234.71	138.21	47.87	362.08
Filtered Reads	Reads Mapped	7,007,348	16,527,538	12,965,972	13,326,759	10,065,430
	Mean Read Length (bp)	98.71	98.86	98.90	98.98	99.00
	Read Length SD (bp)	7.13	6.92	6.85	6.71	6.67
	Mean Reads / bp	27.66	26.13	25.55	24.39	23.31
	Reads/bp SD	723.69	145.95	21.54	19.60	108.90

Table 5 *An. melas* population pair-wise, SNP F_{ST} values per chromosome arm. Regions of heterochromatin in the *An. gambiae* genome were removed from summary statistics. Reprinted from Deitz *et al.* (2016).

	X				2R				2L			
Comparison	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3
West - South	0.019	0.032	0.103	0.055	0.021	0.034	0.070	0.060	0.022	0.034	0.074	0.063
West - Bioko	0.013	0.025	0.074	0.043	0.016	0.028	0.075	0.053	0.016	0.029	0.078	0.058
South - Bioko	0.017	0.029	0.147	0.064	0.018	0.033	0.105	0.086	0.019	0.034	0.115	0.099

	3R				3L				Genome-wide			
Comparison	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3	Q1	Median	Mean	Q3
West - South	0.022	0.034	0.072	0.062	0.021	0.034	0.078	0.064	0.021	0.034	0.075	0.062
West - Bioko	0.016	0.029	0.076	0.059	0.016	0.029	0.076	0.057	0.016	0.028	0.076	0.055
South - Bioko	0.018	0.034	0.111	0.092	0.018	0.034	0.116	0.094	0.018	0.033	0.114	0.091

Table 6 Gene Ontology: Molecular functions for genes harboring significant SNPs found in the bottom 5% Tajima's D regions for the respective populations. Reprinted from Deitz *et al.* (2016).

Molecular Function Category	West - South	West - Bioko	South - Bioko
antioxidant activity (GO:0016209)	0	1	0
binding (GO:0005488)	13	11	32
catalytic activity (GO:0003824)	20	21	38
enzyme regulator activity (GO:0030234)	3	1	6
nucleic acid binding transcription factor activity (GO:0001071)	6	4	8
protein binding transcription factor activity (GO:0000988)	0	0	1
receptor activity (GO:0004872)	4	5	15
structural molecule activity (GO:0005198)	1	0	9
translation regulator activity (GO:0045182)	0	0	2
transporter activity (GO:0005215)	3	1	9
Total Molecular Function Gene Ontology Hits	50	44	120
Genes	64	62	127
SNPs	95	79	188

Table 7 Gene Ontology: Biological processes for genes harboring significant SNPs found in the bottom 5% Tajima's D regions for the respective populations. Reprinted from Deitz *et al.* (2016).

Biological Process Category	West - South	West - Bioko	South - Bioko
apoptotic process (GO:0006915)	2	2	2
biological adhesion (GO:0022610)	1	2	2
biological regulation (GO:0065007)	8	6	17
cellular component organization or biogenesis (GO:0071840)	2	1	4
cellular process (GO:0009987)	13	11	29
developmental process (GO:0032502)	5	5	10
immune system process (GO:0002376)	1	2	3
localization (GO:0051179)	8	6	22
metabolic process (GO:0008152)	30	28	55
multicellular organismal process (GO:0032501)	1	2	8
reproduction (GO:0000003)	1	1	2
response to stimulus (GO:0050896)	2	0	7
Total Biological Process Gene Ontology Hits	74	66	161
Genes	64	62	127
SNPs	95	79	188

Table 8 Gene Ontology: Protein classes for genes harboring significant SNPs found in the bottom 5% Tajima's D regions for the respective populations. Reprinted from Deitz *et al.* (2016).

Protein Class Category	West - South	West - Bioko	South - Bioko
calcium-binding protein (PC00060)	1	0	3
cell adhesion molecule (PC00069)	1	1	0
cell junction protein (PC00070)	0	0	1
cytoskeletal protein (PC00085)	1	0	9
defense/immunity protein (PC00090)	1	2	1
enzyme modulator (PC00095)	2	2	8
extracellular matrix protein (PC00102)	3	2	5
hydrolase (PC00121)	11	10	20
isomerase (PC00135)	0	0	2
kinase (PC00137)	0	1	0
ligase (PC00142)	3	2	4
lyase (PC00144)	0	1	0
membrane traffic protein (PC00150)	0	0	1
nucleic acid binding (PC00171)	7	6	13
oxidoreductase (PC00176)	1	4	4
phosphatase (PC00181)	2	2	2
protease (PC00190)	7	7	17
receptor (PC00197)	4	5	15
signaling molecule (PC00207)	1	1	6
transcription factor (PC00218)	6	4	9
transfer/carrier protein (PC00219)	1	0	2
transferase (PC00220)	1	4	5
transporter (PC00227)	2	1	8
Total Protein Class Gene Ontology Hits	55	55	135
Genes	64	62	127
SNPs	95	79	188

Table 9 Mean Patterson's *D*-statistic values per chromosome arm, resulting from the ABBA-BABA test for introgression using the *An. melas* population tree ((West,Bioko)South)*An. gambiae*). Reprinted from Deitz *et al.* (2016).

Chromosome	*D*-Statistic Mean	*D*-Statistic Std. Error	*D*-Statistic Jackknife Mean	*D*-Statistic Jackknife Std. Error	*D*-Statistic Jackknife Z-Score
X	0.030	0.0048	0.030	0.0048	6.15
2R	0.049	0.0029	0.049	0.0029	16.56
2L	0.021	0.0058	0.021	0.0058	3.69
3R	0.045	0.0030	0.045	0.0030	15.10
3L	0.048	0.0035	0.048	0.0035	13.69
Genome-wide	0.040	0.0018	0.040	0.0018	21.80

Table 10 Number of parental strain reads resulting from the sequencing effort, and the mapped reads and SNPs used to construct the species-specific pseudo-genomes during each iteration. The number of parental strain reads mapped to the species-specific cDNA pseudo-genomes, and the overall mapping efficiency of this step, are also reported.

Species	Sex	Biological Replicate	Reads / Library	Reads Aligned to Spp. Specific Genome	Spp. Total	SNPs	Reads Aligned to AgamP4 Pseudo-Genome	Spp. Total	SNPs	Reads Aligned to AgamP4 cDNA Pseudo-Genome	cDNA Pseudo-Genome Aligned / Library Reads
ARAB	Female	1	59,144,676	51,319,808	200,025,564	616,015	57,941,127	224,760,935	1,393,416	38,550,316	0.65
ARAB	Female	2	58,905,371	51,010,492			57,585,676			37,613,875	0.64
ARAB	Male	1	58,454,004	51,870,320			56,716,992			37,689,738	0.64
ARAB	Male	2	54,069,329	45,824,944			52,517,140			34,858,804	0.64
COLZ	Female	1	57,382,569	46,811,267	183,737,140	1,011,807	56,865,298	218,431,012	785,430	36,675,713	0.64
COLZ	Female	2	57,217,398	48,450,268			56,712,519			36,110,209	0.63
COLZ	Male	1	51,726,503	44,231,501			50,898,565			33,004,947	0.64
COLZ	Male	2	54,682,009	44,244,104			53,954,630			35,149,438	0.64
QUAD	Female	1	63,055,782	53,594,607	197,913,550	574,435	60,228,560	218,694,476	1,393,877	40,079,895	0.64
QUAD	Female	2	53,005,437	46,639,105			50,465,622			33,414,445	0.63
QUAD	Male	1	55,733,151	49,294,593			53,361,875			34,697,831	0.62
QUAD	Male	2	57,084,178	48,385,245			54,638,419			36,509,791	0.64

Table 11 Number of F1 hybrid reads resulting from the sequencing effort. The number of reads mapped to the bi-parental cDNA pseudo-genomes, and the overall mapping efficiency of this step, are also reported.

F1 Hybrid	Sex	Biological Replicate	Reads / Library	Reads Aligned to Bi-Parental AgamP4 cDNA Pseudo- Genome	cDNA Pseudo-Genome Aligned / Library Reads
ARAB x COLZ	Female	1	69,281,988	43,877,381	0.63
ARAB x COLZ	Female	2	65,581,669	41,497,696	0.63
ARAB x COLZ	Male	1	59,705,068	38,184,252	0.64
ARAB x COLZ	Male	2	53,940,601	34,596,662	0.64
COLZ x ARAB	Female	1	63,941,581	39,832,651	0.62
COLZ x ARAB	Female	2	65,834,624	41,186,715	0.63
COLZ x ARAB	Male	1	57,888,755	36,748,874	0.63
COLZ x ARAB	Male	2	57,660,296	36,215,239	0.63
QUAD x COLZ	Female	1	67,356,121	43,647,433	0.65
QUAD x COLZ	Female	2	75,012,804	47,914,798	0.64
QUAD x COLZ	Male	1	67,461,762	44,688,910	0.66
QUAD x COLZ	Male	2	68,687,049	44,601,052	0.65
COLZ x QUAD	Female	1	58,658,067	38,570,792	0.66
COLZ x QUAD	Female	2	248,178,322	162,284,399	0.65
COLZ x QUAD	Male	1	57,430,617	37,452,099	0.65
COLZ x QUAD	Male	2	65,381,139	43,622,785	0.67

Table 12 Number of genes included in the X:A and 2:3 median gene expression ratio analyses at increasing RPKM cut-off thresholds for the *An. coluzzi* - *An. arabiensis* species comparisons.

RPKM Cut-Off (> Value)	X Chromosome	2nd Chromosome	3rd Chromosome	Autosomes
0.0	1,090	6,636	4,513	11,149
0.2	952	5,852	4,005	9,857
0.5	851	5,305	3,625	8,930
1.0	735	4,752	3,246	7,998
2.0	594	4,055	2,708	6,763
5.0	402	2,836	1,848	4,684
10.0	248	1,872	1,222	3,094

Table 13 Median X:A expression ratios, and their 95% confidence intervals (CI), for each sample of the *An. coluzzii* - *An. arabiensis* species comparison at increasing minimum RPKM cut-off thresholds. The results of the between sample ANOVA (F-value and probability > F) and Kruskal-Wallis test (X^2 and p-value) are reported for each minimum RPKM cut-off threshold. Table continued on next page.

RPKM Cut-Off			> 0.0		>0.2		>0.5	
Species / F1 Hybrid	Sex	Biological Replicate	Median	95% CI	Median	95% CI	Median	95% CI
ARAB	Female	1	0.89	(0.76,1.02)	0.91	(0.82,1.01)	0.89	(0.78,0.98)
ARAB	Female	2	0.81	(0.73,0.9)	0.81	(0.74,0.88)	0.81	(0.73,0.88)
ARAB	Male	1	0.75	(0.66,0.83)	0.78	(0.71,0.86)	0.79	(0.72,0.87)
ARAB	Male	2	0.75	(0.66,0.83)	0.78	(0.69,0.86)	0.78	(0.72,0.84)
COLZ	Female	1	0.87	(0.78,0.97)	0.87	(0.79,0.95)	0.88	(0.8,0.98)
COLZ	Female	2	0.91	(0.82,0.99)	0.92	(0.83,1.01)	0.92	(0.82,1)
COLZ	Male	1	0.80	(0.72,0.89)	0.80	(0.73,0.87)	0.81	(0.72,0.89)
COLZ	Male	2	0.82	(0.75,0.9)	0.82	(0.73,0.89)	0.84	(0.77,0.92)
ARAB x COLZ	Female	1	0.85	(0.76,0.96)	0.87	(0.78,0.95)	0.87	(0.76,0.97)
ARAB x COLZ	Female	2	0.82	(0.72,0.9)	0.84	(0.75,0.94)	0.88	(0.78,0.98)
ARAB x COLZ	Male	1	0.74	(0.62,0.85)	0.83	(0.73,0.97)	0.79	(0.71,0.86)
ARAB x COLZ	Male	2	0.71	(0.62,0.78)	0.76	(0.68,0.85)	0.77	(0.71,0.84)
COLZ x ARAB	Female	1	0.74	(0.67,0.81)	0.79	(0.69,0.9)	0.83	(0.73,0.92)
COLZ x ARAB	Female	2	0.76	(0.68,0.84)	0.83	(0.74,0.93)	0.85	(0.75,0.94)
COLZ x ARAB	Male	1	0.77	(0.69,0.85)	0.79	(0.71,0.86)	0.80	(0.71,0.88)
COLZ x ARAB	Male	2	0.78	(0.71,0.84)	0.79	(0.71,0.87)	0.80	(0.71,0.88)
ANOVA F-value			3.13		2.37		1.89	
ANOVA Pr(>F)			0.000		0.002		0.020	
Kruskal-Wallis X^2			26.17		22.93		21.92	
Kruskal-Wallis p-value			0.036		0.086		0.110	

Table 13, continued.

RPKM Cut-Off			>1.0		>2.0		>5.0		>10.0	
Species / F1 Hybrid	Sex	Biological Replicate	Median	95% CI	Median	95% CI	Median	95% CI	Median	95% CI
ARAB	Female	1	0.97	(0.85,1.1)	0.99	(0.87,1.1)	0.94	(0.81,1.03)	1.13	(0.93,1.35)
ARAB	Female	2	0.84	(0.77,0.9)	0.86	(0.76,0.96)	0.94	(0.83,1.05)	0.97	(0.79,1.13)
ARAB	Male	1	0.82	(0.74,0.9)	0.83	(0.73,0.91)	0.90	(0.75,1.04)	0.99	(0.79,1.14)
ARAB	Male	2	0.77	(0.7,0.83)	0.82	(0.72,0.91)	0.84	(0.67,0.96)	0.96	(0.75,1.13)
COLZ	Female	1	0.91	(0.84,0.98)	0.88	(0.78,0.97)	0.90	(0.78,1.05)	1.00	(0.77,1.19)
COLZ	Female	2	0.93	(0.86,1)	0.93	(0.84,1.02)	0.92	(0.81,1.06)	1.07	(0.87,1.33)
COLZ	Male	1	0.81	(0.74,0.86)	0.82	(0.73,0.93)	0.88	(0.75,1.04)	1.05	(0.86,1.3)
COLZ	Male	2	0.84	(0.77,0.91)	0.81	(0.73,0.87)	0.86	(0.72,1.03)	0.95	(0.7,1.13)
ARAB x COLZ	Female	1	0.93	(0.86,1.02)	0.96	(0.87,1.05)	0.99	(0.85,1.13)	1.03	(0.87,1.19)
ARAB x COLZ	Female	2	0.93	(0.85,1.02)	0.98	(0.89,1.09)	0.98	(0.86,1.11)	1.04	(0.9,1.18)
ARAB x COLZ	Male	1	0.83	(0.72,0.95)	0.91	(0.77,1.04)	0.94	(0.83,1.06)	0.94	(0.67,1.17)
ARAB x COLZ	Male	2	0.79	(0.71,0.87)	0.86	(0.73,0.96)	0.84	(0.67,0.96)	0.92	(0.77,1.06)
COLZ x ARAB	Female	1	0.85	(0.75,0.93)	0.90	(0.79,1.01)	0.97	(0.86,1.08)	0.98	(0.84,1.1)
COLZ x ARAB	Female	2	0.90	(0.82,0.99)	0.95	(0.81,1.06)	0.96	(0.83,1.05)	1.00	(0.83,1.15)
COLZ x ARAB	Male	1	0.87	(0.77,0.97)	0.88	(0.81,0.97)	0.91	(0.75,1.08)	0.99	(0.82,1.16)
COLZ x ARAB	Male	2	0.87	(0.78,0.98)	0.87	(0.79,0.95)	0.91	(0.76,1.05)	0.97	(0.8,1.15)
ANOVA F-value			1.38		0.88		0.53		0.38	
ANOVA Pr(>F)			0.146		0.589		0.925		0.984	
Kruskal-Wallis X^2			19.37		20.45		13.92		5.34	
Kruskal-Wallis p-value			0.198		0.155		0.531		0.989	

Table 14 Median 2:3 expression ratios, and their 95% confidence intervals (CI), for each sample of the *An. coluzzii* - *An. arabiensis* species comparison at increasing minimum RPKM cut-off thresholds. The results of the between sample ANOVA (F-value and probability > F) and Kruskal-Wallis test (X^2 and p-value) are reported for each minimum RPKM cut-off threshold. Table continued on next page.

RPKM Cut-Off			> 0.0		>0.2		>0.5	
Species / F1 Hybrid	Sex	Biological Replicate	Median	95% CI	Median	95% CI	Median	95% CI
ARAB	Female	1	1.14	(1.07,1.21)	1.17	(1.11,1.23)	1.14	(1.07,1.2)
ARAB	Female	2	1.06	(1.01,1.1)	1.05	(1.01,1.09)	1.06	(1.01,1.11)
ARAB	Male	1	1.06	(1.02,1.11)	1.05	(1.01,1.09)	1.04	(1,1.08)
ARAB	Male	2	1.06	(1.01,1.1)	1.05	(1.01,1.09)	1.06	(1,1.1)
COLZ	Female	1	1.06	(1.01,1.1)	1.03	(0.99,1.08)	1.06	(1.01,1.11)
COLZ	Female	2	1.07	(1.03,1.11)	1.06	(1.02,1.11)	1.06	(1.01,1.1)
COLZ	Male	1	1.09	(1.04,1.14)	1.09	(1.04,1.13)	1.08	(1.03,1.12)
COLZ	Male	2	1.09	(1.05,1.14)	1.06	(1.01,1.11)	1.04	(0.99,1.08)
ARAB x COLZ	Female	1	1.03	(0.98,1.08)	1.05	(1.01,1.1)	1.08	(1.03,1.12)
ARAB x COLZ	Female	2	1.03	(0.98,1.08)	1.05	(1,1.1)	1.07	(1.02,1.11)
ARAB x COLZ	Male	1	1.15	(1.07,1.22)	1.18	(1.11,1.25)	1.15	(1.08,1.22)
ARAB x COLZ	Male	2	1.07	(1.02,1.11)	1.07	(1.02,1.11)	1.06	(1.01,1.1)
COLZ x ARAB	Female	1	1.02	(0.98,1.06)	1.04	(1,1.08)	1.07	(1.03,1.12)
COLZ x ARAB	Female	2	1.05	(1,1.09)	1.06	(1.02,1.11)	1.09	(1.04,1.13)
COLZ x ARAB	Male	1	1.06	(1.01,1.11)	1.05	(1.01,1.1)	1.10	(1.05,1.15)
COLZ x ARAB	Male	2	1.06	(1.01,1.11)	1.04	(1,1.08)	1.06	(1.01,1.11)
ANOVA F-value			22.99		19.45		14.24	
ANOVA Pr(>F)			<2e-16		<2e-16		<2e-16	
Kruskal-Wallis X^2			76.46		44.18		18.87	
Kruskal-Wallis p-value			0.000		0.000		0.220	

Table 14, continued.

RPKM Cut-Off			>1.0		>2.0		>5.0		>10.0	
Species / F1 Hybrid	Sex	Biological Replicate	Median	95% CI	Median	95% CI	Median	95% CI	Median	95% CI
ARAB	Female	1	1.14	(1.08,1.19)	1.05	(0.99,1.1)	0.99	(0.94,1.04)	0.98	(0.93,1.04)
ARAB	Female	2	1.05	(1.01,1.09)	0.97	(0.93,1.01)	0.95	(0.91,1.01)	0.96	(0.91,1.01)
ARAB	Male	1	1.00	(0.97,1.04)	0.98	(0.93,1.01)	0.94	(0.9,0.99)	0.95	(0.91,1)
ARAB	Male	2	1.05	(1.01,1.09)	0.99	(0.95,1.03)	0.96	(0.92,1.01)	0.96	(0.92,1)
COLZ	Female	1	1.03	(0.98,1.07)	0.98	(0.94,1.02)	0.97	(0.92,1.03)	0.95	(0.9,1)
COLZ	Female	2	1.04	(0.99,1.08)	0.97	(0.92,1.02)	0.96	(0.91,1.01)	0.95	(0.89,1.01)
COLZ	Male	1	1.07	(1.02,1.11)	0.97	(0.93,1.01)	0.97	(0.92,1.03)	0.97	(0.93,1.01)
COLZ	Male	2	1.02	(0.98,1.06)	0.99	(0.94,1.03)	0.96	(0.91,1.01)	0.97	(0.91,1.03)
ARAB x COLZ	Female	1	1.08	(1.03,1.13)	1.02	(0.98,1.06)	1.01	(0.96,1.07)	1.03	(0.97,1.09)
ARAB x COLZ	Female	2	1.09	(1.04,1.13)	1.03	(0.99,1.07)	1.03	(0.97,1.09)	1.01	(0.96,1.06)
ARAB x COLZ	Male	1	1.14	(1.08,1.2)	1.13	(1.08,1.19)	1.04	(0.98,1.1)	0.97	(0.91,1.03)
ARAB x COLZ	Male	2	1.06	(1.02,1.11)	1.00	(0.96,1.04)	1.00	(0.94,1.05)	1.05	(0.99,1.11)
COLZ x ARAB	Female	1	1.09	(1.04,1.15)	1.04	(0.99,1.08)	0.98	(0.92,1.04)	1.01	(0.97,1.07)
COLZ x ARAB	Female	2	1.11	(1.06,1.15)	1.05	(1,1.1)	0.98	(0.93,1.04)	1.02	(0.97,1.08)
COLZ x ARAB	Male	1	1.10	(1.06,1.14)	1.02	(0.98,1.06)	0.98	(0.93,1.03)	1.00	(0.95,1.04)
COLZ x ARAB	Male	2	1.09	(1.04,1.13)	1.03	(0.99,1.07)	0.97	(0.92,1.01)	1.00	(0.94,1.05)
ANOVA F-value			11.79		9.25		5.47		2.29	
ANOVA Pr(>F)			<2e-16		<2e-16		0.000		0.003	
Kruskal-Wallis X^2			44.71		56.14		44.94		11.05	
Kruskal-Wallis p-value			0.000		0.000		0.000		0.749	

Table 15 The results of a Tukey's post-hoc test for pair-wise comparisons within species or hybrids of the *An. coluzzii* - *An. arabiensis* species comparison. P-values are reported for X:A and 2:3 expression ratios for increasing RPKM cut-offs. Significant p-values (<0.05) are shown in bold font and indicate significant differences in the mean of the distributions of the two samples in question (spp1 sex 1, spp2 sex 2).

				RPKM > 0.0		RPKM > 0.2		RPKM > 0.5		RPKM > 1.0		RPKM > 2.0		RPKM > 5.0		RPKM > 10.0	
spp1	sex1	spp2	sex2	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3
ARAB	female2	ARAB	female1	**0.04**	**0.00**	0.12	**0.00**	0.20	**0.00**	0.52	**0.00**	0.85	**0.00**	0.98	**0.00**	1.00	0.10
ARAB	male1	ARAB	female1	0.09	**0.00**	0.23	**0.00**	0.39	**0.00**	0.77	**0.00**	0.97	**0.00**	1.00	**0.02**	1.00	0.68
ARAB	male1	ARAB	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB	male2	ARAB	female1	**0.04**	**0.00**	0.12	**0.00**	0.22	**0.00**	0.54	**0.00**	0.88	**0.00**	0.99	**0.00**	1.00	0.20
ARAB	male2	ARAB	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB	male2	ARAB	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB x COLZ	female2	ARAB x COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB x COLZ	male1	ARAB x COLZ	female1	**0.01**	**0.00**	0.07	**0.00**	0.22	**0.00**	0.48	**0.00**	0.84	**0.00**	0.99	**0.00**	1.00	**0.04**
ARAB x COLZ	male1	ARAB x COLZ	female2	**0.01**	**0.00**	0.06	**0.00**	0.19	**0.00**	0.46	**0.00**	0.82	**0.00**	0.98	**0.00**	1.00	**0.05**
ARAB x COLZ	male2	ARAB x COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB x COLZ	male2	ARAB x COLZ	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB x COLZ	male2	ARAB x COLZ	male1	**0.00**	**0.00**	**0.04**	**0.00**	0.15	**0.00**	0.39	**0.00**	0.79	**0.00**	0.99	**0.00**	1.00	0.39
COLZ	female2	COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male1	COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male1	COLZ	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male2	COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male2	COLZ	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male2	COLZ	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	female2	COLZ x ARAB	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	male1	COLZ x ARAB	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	male1	COLZ x ARAB	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	male2	COLZ x ARAB	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	male2	COLZ x ARAB	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	male2	COLZ x ARAB	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 16 The results of a Dunn's post-hoc test for pair-wise comparisons within species or hybrids of the *An. coluzzii* - *An. arabiensis* species comparison. P-values are reported for X:A and 2:3 expression ratios for increasing RPKM cut-offs. Significant p-values (<0.05) are shown in bold font and indicate significant differences in the mean of the distributions of the two samples in question (spp1 sex 1, spp2 sex 2).

				RPKM > 0.0		RPKM > 0.2		RPKM > 0.5		RPKM > 1.0		RPKM > 2.0		RPKM > 5.0		RPKM > 10.0	
spp1	sex1	spp2	sex2	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3
ARAB	female1	ARAB	female2	1.00	1.00	1.00	0.70	1.00	1.00	1.00	0.11	1.00	1.00	1.00	1.00	1.00	1.00
ARAB	female1	ARAB	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**0.00**	1.00	1.00	1.00	1.00	1.00	1.00
ARAB	female1	ARAB	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	0.11	1.00	1.00	1.00	1.00	1.00	1.00
ARAB	female2	ARAB	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB	female2	ARAB	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB	male1	ARAB	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB x COLZ	female1	ARAB x COLZ	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB x COLZ	female1	ARAB x COLZ	male1	1.00	**0.00**	1.00	**0.01**	1.00	1.00	1.00	1.00	1.00	0.81	1.00	1.00	1.00	1.00
ARAB x COLZ	female1	ARAB x COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB x COLZ	female2	ARAB x COLZ	male1	1.00	**0.00**	1.00	**0.02**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB x COLZ	female2	ARAB x COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
ARAB x COLZ	male1	ARAB x COLZ	male2	1.00	0.27	1.00	0.10	1.00	1.00	1.00	1.00	1.00	**0.03**	1.00	1.00	1.00	1.00
COLZ	female1	COLZ	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	female1	COLZ	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	female1	COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	female2	COLZ	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	female2	COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male1	COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	female1	COLZ x ARAB	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	female1	COLZ x ARAB	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	female1	COLZ x ARAB	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	female2	COLZ x ARAB	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	female2	COLZ x ARAB	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x ARAB	male1	COLZ x ARAB	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 17 Results of an ANOVA test (F-value and p-values, Pr>(F)) for differences between the mean of the X:A and 2:3 expression ratio distributions within each sample of the the *An. coluzzii* - *An. arabiensis* species comparison at increasing minimum RPKM cut-off thresholds. Significant p-values (≤0.05) are shown in bold font.

RPKM Cut-Off			> 0.0		> 0.2		> 0.5		> 1.0		> 2.0		> 5.0		> 10.0	
Species / F1 Hybrid	Sex	Biological Replicate	F-value	Pr(>F)	F-value	Pr(>F)	F-value	Pr(>F)	F-value	Pr(>F)	F-value	Pr(>F)	F-value	Pr(>F)	F-value	Pr(>F)
ARAB	Female	1	5.16	**0.023**	5.12	**0.024**	3.99	**0.046**	3.83	**0.050**	2.10	0.147	1.38	0.240	0.76	0.383
ARAB	Female	2	5.08	**0.024**	4.43	**0.035**	3.85	**0.050**	2.94	0.086	1.30	0.254	0.81	0.369	0.47	0.492
ARAB	Male	1	4.38	**0.036**	3.99	**0.046**	3.19	0.074	2.18	0.140	1.23	0.267	0.67	0.413	0.28	0.597
ARAB	Male	2	4.92	**0.027**	4.39	**0.036**	3.79	0.052	2.97	0.085	1.61	0.205	0.99	0.320	0.45	0.504
COLZ	Female	1	2.49	0.115	1.98	0.160	1.74	0.187	0.86	0.355	0.19	0.665	0.03	0.854	0.01	0.915
COLZ	Female	2	2.29	0.130	1.88	0.170	1.53	0.216	0.79	0.374	0.15	0.694	0.03	0.864	0.02	0.897
COLZ	Male	1	3.86	**0.050**	3.44	0.064	2.92	0.087	2.03	0.154	0.81	0.369	0.49	0.486	0.13	0.720
COLZ	Male	2	3.09	0.079	2.37	0.124	1.82	0.177	1.02	0.312	0.37	0.541	0.13	0.721	0.01	0.914
ARAB x COLZ	Female	1	2.66	0.103	2.79	0.095	2.67	0.102	2.14	0.144	0.92	0.338	0.60	0.439	0.35	0.554
ARAB x COLZ	Female	2	2.92	0.088	2.88	0.090	2.76	0.097	2.31	0.129	1.10	0.294	0.78	0.377	0.33	0.563
ARAB x COLZ	Male	1	4.87	**0.027**	4.83	**0.028**	4.22	**0.040**	3.58	0.059	2.82	0.094	1.85	0.174	0.83	0.362
ARAB x COLZ	Male	2	4.28	**0.039**	3.85	**0.050**	3.50	0.062	2.78	0.095	1.27	0.260	0.85	0.357	0.67	0.412
COLZ x ARAB	Female	1	2.50	0.114	2.49	0.115	2.45	0.117	2.08	0.150	1.09	0.297	0.52	0.469	0.35	0.553
COLZ x ARAB	Female	2	2.90	0.089	2.78	0.095	2.80	0.094	2.44	0.118	1.17	0.280	0.56	0.453	0.39	0.531
COLZ x ARAB	Male	1	3.16	0.076	2.89	0.089	2.99	0.084	2.43	0.119	1.15	0.284	0.58	0.448	0.30	0.582
COLZ x ARAB	Male	2	3.09	0.079	2.78	0.095	2.71	0.100	2.33	0.127	1.27	0.261	0.58	0.448	0.35	0.555

Table 18 Results of the Kruskal-Wallis test (X^2 and p-values) for differences between the median of the X:A and 2:3 expression ratio distributions within each sample of the the *An. coluzzii* - *An. arabiensis* species comparison at increasing minimum RPKM cut-off thresholds. Significant p-values (≤0.05) are shown in bold font.

RPKM Cut-Off			> 0.0		> 0.2		> 0.5		> 1.0		> 2.0		> 5.0		> 10.0	
Species / F1 Hybrid	Sex	Biological Replicate	X^2	p-value	X^2	p-value	X^2	p-value	X^2	p-value	X^2	p-value	X^2	p-value	X^2	p-value
ARAB	Female	1	15.62	**0.000**	18.48	**0.000**	13.67	**0.000**	8.46	**0.004**	0.92	0.337	0.00	0.958	1.76	0.184
ARAB	Female	2	29.69	**0.000**	28.54	**0.000**	28.04	**0.000**	14.62	**0.000**	2.39	0.122	0.20	0.654	0.08	0.772
ARAB	Male	1	38.87	**0.000**	37.27	**0.000**	31.26	**0.000**	13.02	**0.000**	4.73	**0.030**	0.48	0.490	0.37	0.543
ARAB	Male	2	46.04	**0.000**	43.45	**0.000**	39.51	**0.000**	23.32	**0.000**	8.93	**0.003**	2.41	0.120	0.03	0.874
ARAB x COLZ	Female	1	9.47	**0.002**	12.88	**0.000**	14.25	**0.000**	8.39	**0.004**	0.01	0.908	0.02	0.888	0.02	0.899
ARAB x COLZ	Female	2	12.68	**0.000**	15.34	**0.000**	16.69	**0.000**	11.36	**0.001**	0.47	0.493	0.09	0.765	0.02	0.879
ARAB x COLZ	Male	1	42.59	**0.000**	42.99	**0.000**	34.10	**0.000**	21.91	**0.000**	11.44	**0.001**	4.91	**0.027**	0.00	0.989
ARAB x COLZ	Male	2	51.17	**0.000**	46.80	**0.000**	43.76	**0.000**	28.44	**0.000**	6.30	**0.012**	3.78	0.052	2.00	0.158
COLZ	Female	1	15.56	**0.000**	14.19	**0.000**	14.60	**0.000**	4.80	**0.028**	0.51	0.477	0.00	0.946	0.62	0.433
COLZ	Female	2	9.96	**0.002**	9.39	**0.002**	8.43	**0.004**	2.14	0.144	0.01	0.922	0.23	0.631	1.41	0.235
COLZ	Male	1	39.38	**0.000**	38.69	**0.000**	34.59	**0.000**	20.53	**0.000**	6.65	**0.010**	3.37	0.067	0.04	0.843
COLZ	Male	2	32.79	**0.000**	26.79	**0.000**	21.81	**0.000**	10.55	**0.001**	4.20	**0.040**	1.56	0.212	0.00	0.979
COLZ x ARAB	Female	1	19.81	**0.000**	24.32	**0.000**	27.46	**0.000**	20.64	**0.000**	4.56	**0.033**	0.20	0.651	0.02	0.894
COLZ x ARAB	Female	2	19.39	**0.000**	20.52	**0.000**	24.09	**0.000**	17.14	**0.000**	2.35	0.125	0.00	0.970	0.04	0.849
COLZ x ARAB	Male	1	31.44	**0.000**	32.30	**0.000**	38.61	**0.000**	27.38	**0.000**	8.03	**0.005**	2.27	0.132	0.48	0.487
COLZ x ARAB	Male	2	31.85	**0.000**	31.95	**0.000**	34.99	**0.000**	27.61	**0.000**	10.04	**0.002**	2.08	0.149	0.64	0.425

Table 19 Number of genes included in the X:A and 2:3 median gene expression ratio analyses at increasing RPKM cut-off thresholds for the *An. coluzzi* - *An. quadriannulatus* species comparisons.

RPKM Cut-Off (> Value)	X Chromosome	2nd Chromosome	3rd Chromosome	Autosomes
0.0	1,118	6,736	4,523	11,259
0.2	982	5,967	4,014	9,981
0.5	875	5,442	3,658	9,100
1.0	763	4,895	3,261	8,156
2.0	621	4,118	2,707	6,825
5.0	403	2,759	1,788	4,547
10.0	230	1,733	1,129	2,862

Table 20 Median X:A expression ratios, and their 95% confidence intervals (CI), for each sample of the *An. coluzzii* - *An. quadriannulatus* species comparison at increasing minimum RPKM cut-off thresholds. The results of the between sample ANOVA (F-value and probability > F) and Kruskal-Wallis test (X^2 and p-value) are reported for each minimum RPKM cut-off threshold. Table continued on next page.

RPKM Cut-Off			> 0.0		>0.2		>0.5	
Species / F1 Hybrid	Sex	Biological Replicate	Median	95% CI	Median	95% CI	Median	95% CI
QUAD	Female	1	0.82	(0.72,0.92)	0.81	(0.7,0.91)	0.84	(0.74,0.93)
QUAD	Female	2	0.84	(0.7,0.97)	0.84	(0.75,0.95)	0.86	(0.77,0.96)
QUAD	Male	1	0.70	(0.62,0.76)	0.71	(0.64,0.78)	0.70	(0.64,0.75)
QUAD	Male	2	0.74	(0.63,0.84)	0.76	(0.68,0.86)	0.75	(0.66,0.84)
COLZ	Female	1	0.87	(0.78,0.96)	0.85	(0.76,0.92)	0.88	(0.79,0.97)
COLZ	Female	2	0.90	(0.82,0.99)	0.89	(0.81,0.98)	0.90	(0.8,0.98)
COLZ	Male	1	0.79	(0.7,0.87)	0.79	(0.71,0.86)	0.79	(0.71,0.87)
COLZ	Male	2	0.82	(0.76,0.9)	0.79	(0.71,0.86)	0.81	(0.73,0.88)
QUAD x COLZ	Female	1	0.78	(0.69,0.89)	0.79	(0.7,0.86)	0.82	(0.72,0.9)
QUAD x COLZ	Female	2	0.77	(0.67,0.86)	0.79	(0.71,0.87)	0.82	(0.73,0.9)
QUAD x COLZ	Male	1	0.70	(0.63,0.78)	0.72	(0.64,0.8)	0.78	(0.71,0.85)
QUAD x COLZ	Male	2	0.71	(0.65,0.79)	0.70	(0.62,0.76)	0.74	(0.65,0.83)
COLZ x QUAD	Female	1	0.92	(0.83,1.01)	0.91	(0.81,1)	0.95	(0.86,1.03)
COLZ x QUAD	Female	2	0.90	(0.81,0.99)	0.90	(0.81,0.99)	0.92	(0.82,1)
COLZ x QUAD	Male	1	0.77	(0.69,0.84)	0.79	(0.71,0.88)	0.82	(0.73,0.9)
COLZ x QUAD	Male	2	0.79	(0.7,0.88)	0.80	(0.7,0.88)	0.85	(0.77,0.94)
ANOVA F-value			1.56		1.43		1.41	
ANOVA Pr(>F)			0.08		0.12		0.13	
Kruskal-Wallis X^2			44.96		44.98		51.16	
Kruskal-Wallis p-value			0.00		0.00		0.00	

Table 20, continued.

RPKM Cut-Off			>1.0		>2.0		>5.0		>10.0	
Species / F1 Hybrid	Sex	Biological Replicate	Median	95% CI	Median	95% CI	Median	95% CI	Median	95% CI
QUAD	Female	1	0.88	(0.81,0.97)	0.85	(0.74,0.95)	0.82	(0.69,0.95)	0.99	(0.81,1.17)
QUAD	Female	2	0.85	(0.74,0.96)	0.85	(0.74,0.95)	0.87	(0.75,0.99)	1.09	(0.91,1.26)
QUAD	Male	1	0.75	(0.68,0.85)	0.77	(0.67,0.84)	0.80	(0.66,0.92)	0.89	(0.77,1.02)
QUAD	Male	2	0.79	(0.71,0.87)	0.81	(0.69,0.92)	0.79	(0.68,0.9)	0.98	(0.75,1.18)
COLZ	Female	1	0.88	(0.82,0.94)	0.85	(0.74,0.94)	0.88	(0.77,0.98)	1.06	(0.83,1.27)
COLZ	Female	2	0.91	(0.85,0.98)	0.89	(0.79,0.97)	0.90	(0.78,1)	1.06	(0.82,1.24)
COLZ	Male	1	0.79	(0.74,0.85)	0.80	(0.72,0.9)	0.87	(0.74,1)	1.04	(0.85,1.21)
COLZ	Male	2	0.81	(0.75,0.87)	0.79	(0.74,0.85)	0.86	(0.72,0.97)	1.04	(0.85,1.27)
QUAD x COLZ	Female	1	0.89	(0.8,1)	0.90	(0.81,0.98)	0.96	(0.87,1.08)	1.03	(0.83,1.21)
QUAD x COLZ	Female	2	0.90	(0.81,1)	0.92	(0.83,1.01)	0.97	(0.84,1.11)	0.97	(0.81,1.1)
QUAD x COLZ	Male	1	0.79	(0.71,0.86)	0.82	(0.73,0.9)	0.88	(0.73,0.99)	0.98	(0.81,1.16)
QUAD x COLZ	Male	2	0.80	(0.73,0.89)	0.83	(0.73,0.92)	0.88	(0.74,1.02)	1.01	(0.85,1.2)
COLZ x QUAD	Female	1	1.00	(0.9,1.12)	1.02	(0.92,1.11)	1.04	(0.91,1.15)	1.03	(0.72,1.22)
COLZ x QUAD	Female	2	0.95	(0.87,1.04)	0.99	(0.9,1.08)	1.04	(0.92,1.15)	1.06	(0.76,1.29)
COLZ x QUAD	Male	1	0.86	(0.79,0.94)	0.88	(0.8,0.96)	0.95	(0.83,1.08)	1.04	(0.86,1.18)
COLZ x QUAD	Male	2	0.88	(0.79,0.97)	0.90	(0.79,0.99)	0.96	(0.83,1.09)	1.03	(0.79,1.23)
ANOVA F-value			1.39		1.09		1.13		0.92	
ANOVA Pr(>F)			0.14		0.36		0.32		0.54	
Kruskal-Wallis X^2			53.85		53.11		43.97		20.46	
Kruskal-Wallis p-value			0.00		0.00		0.00		0.16	

Table 21 Median 2:3 expression ratios, and their 95% confidence intervals (CI), for each sample of the *An. coluzzii* - *An. quadriannulatus* species comparison at increasing minimum RPKM cut-off thresholds. The results of the between sample ANOVA (F-value and probability > F) and Kruskal-Wallis test (X^2 and p-value) are reported for each minimum RPKM cut-off threshold. Table continued on next page.

RPKM Cut-Off			> 0.0		>0.2		>0.5	
Species / F1 Hybrid	Sex	Biological Replicate	Median	95% CI	Median	95% CI	Median	95% CI
QUAD	Female	1	1.12	(1.05,1.19)	1.13	(1.08,1.19)	1.10	(1.04,1.16)
QUAD	Female	2	1.13	(1.06,1.2)	1.13	(1.08,1.19)	1.11	(1.06,1.17)
QUAD	Male	1	1.03	(0.99,1.07)	1.03	(1,1.07)	1.04	(1,1.09)
QUAD	Male	2	1.15	(1.09,1.21)	1.17	(1.11,1.23)	1.12	(1.07,1.17)
COLZ	Female	1	1.05	(1,1.09)	1.03	(0.98,1.08)	1.04	(0.99,1.09)
COLZ	Female	2	1.06	(1.02,1.1)	1.06	(1.01,1.1)	1.04	(0.99,1.09)
COLZ	Male	1	1.08	(1.04,1.13)	1.09	(1.04,1.13)	1.07	(1.03,1.12)
COLZ	Male	2	1.08	(1.04,1.12)	1.06	(1.01,1.1)	1.03	(0.99,1.08)
QUAD x COLZ	Female	1	1.05	(1.01,1.09)	1.06	(1.02,1.1)	1.06	(1.01,1.11)
QUAD x COLZ	Female	2	1.07	(1.02,1.11)	1.06	(1,1.1)	1.07	(1.02,1.12)
QUAD x COLZ	Male	1	1.05	(1,1.1)	1.06	(1.02,1.11)	1.05	(1.01,1.1)
QUAD x COLZ	Male	2	1.07	(1.03,1.12)	1.08	(1.03,1.13)	1.07	(1.03,1.12)
COLZ x QUAD	Female	1	1.05	(1,1.1)	1.04	(1,1.07)	1.01	(0.97,1.05)
COLZ x QUAD	Female	2	1.04	(0.99,1.09)	1.06	(1.02,1.11)	1.04	(0.99,1.09)
COLZ x QUAD	Male	1	1.04	(0.99,1.09)	1.04	(1,1.08)	1.04	(1,1.08)
COLZ x QUAD	Male	2	1.05	(1,1.09)	1.04	(1,1.08)	1.03	(0.99,1.07)
ANOVA F-value			8.31		7.15		6.07	
ANOVA Pr(>F)			0.00		0.00		0.00	
Kruskal-Wallis X^2			36.97		30.87		15.64	
Kruskal-Wallis p-value			0.00		0.01		0.41	

Table 21, continued.

RPKM Cut-Off			>1.0		>2.0		>5.0		>10.0	
Species / F1 Hybrid	Sex	Biological Replicate	Median	95% CI	Median	95% CI	Median	95% CI	Median	95% CI
QUAD	Female	1	1.11	(1.04,1.16)	1.08	(1.03,1.13)	1.01	(0.94,1.08)	0.96	(0.89,1.01)
QUAD	Female	2	1.09	(1.03,1.15)	1.10	(1.04,1.16)	1.02	(0.97,1.08)	0.98	(0.92,1.04)
QUAD	Male	1	1.06	(1.01,1.11)	1.02	(0.99,1.06)	0.98	(0.94,1.03)	0.99	(0.92,1.04)
QUAD	Male	2	1.10	(1.04,1.17)	1.11	(1.05,1.17)	1.01	(0.95,1.08)	0.93	(0.86,1)
COLZ	Female	1	1.01	(0.96,1.06)	0.98	(0.94,1.02)	0.97	(0.92,1.03)	0.98	(0.93,1.03)
COLZ	Female	2	1.02	(0.97,1.06)	0.97	(0.93,1.02)	0.99	(0.93,1.04)	1.00	(0.94,1.07)
COLZ	Male	1	1.05	(1,1.09)	0.97	(0.94,1.01)	0.99	(0.93,1.04)	0.97	(0.9,1.03)
COLZ	Male	2	1.00	(0.96,1.04)	0.98	(0.94,1.01)	0.97	(0.93,1.01)	0.99	(0.94,1.03)
QUAD x COLZ	Female	1	1.06	(1.02,1.1)	1.00	(0.96,1.05)	1.00	(0.95,1.06)	1.02	(0.97,1.09)
QUAD x COLZ	Female	2	1.08	(1.04,1.12)	1.01	(0.98,1.05)	1.01	(0.95,1.07)	1.02	(0.96,1.06)
QUAD x COLZ	Male	1	1.04	(1,1.08)	1.00	(0.96,1.03)	0.98	(0.93,1.03)	0.99	(0.94,1.04)
QUAD x COLZ	Male	2	1.04	(0.99,1.08)	1.01	(0.97,1.05)	0.99	(0.94,1.04)	1.01	(0.96,1.05)
COLZ x QUAD	Female	1	1.02	(0.96,1.07)	1.00	(0.95,1.04)	1.00	(0.94,1.06)	1.00	(0.95,1.05)
COLZ x QUAD	Female	2	1.01	(0.96,1.05)	1.00	(0.95,1.04)	0.98	(0.91,1.03)	0.99	(0.94,1.06)
COLZ x QUAD	Male	1	1.04	(0.99,1.08)	1.00	(0.96,1.04)	0.97	(0.92,1.01)	0.99	(0.93,1.04)
COLZ x QUAD	Male	2	1.02	(0.98,1.06)	1.00	(0.96,1.04)	0.96	(0.91,1)	1.01	(0.96,1.07)
ANOVA F-value			5.15		4.84		2.94		1.83	
ANOVA Pr(>F)			0.00		0.00		0.00		0.03	
Kruskal-Wallis X^2			18.25		32.06		11.55		31.55	
Kruskal-Wallis p-value			0.25		0.01		0.71		0.01	

Table 22 The results of a Tukey's post-hoc test for pair-wise comparisons within species or hybrids of the *An. coluzzii* - *An. quandriannulatus* species comparison. P-values are reported for X:A and 2:3 expression ratios for increasing RPKM cut-offs. Significant p-values (<0.05) are shown in bold font and indicate significant differences in the mean of the distributions of the two samples in question (spp1 sex 1, spp2 sex 2).

				RPKM > 0.0		RPKM > 0.2		RPKM > 0.5		RPKM > 1.0		RPKM > 2.0		RPKM > 5.0		RPKM > 10.0	
spp1	sex1	spp2	sex2	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3
COLZ	female2	COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male1	COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male1	COLZ	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male2	COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male2	COLZ	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male2	COLZ	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	female2	COLZ x QUAD	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	male1	COLZ x QUAD	female1	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	male1	COLZ x QUAD	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	male2	COLZ x QUAD	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	male2	COLZ x QUAD	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	male2	COLZ x QUAD	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD	female2	QUAD	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD	male1	QUAD	female1	0.59	**0.00**	0.81	**0.00**	0.91	**0.00**	0.96	**0.00**	0.99	**0.01**	1.00	0.30	1.00	0.77
QUAD	male1	QUAD	female2	0.32	**0.00**	0.62	**0.00**	0.71	**0.00**	0.84	**0.00**	0.95	**0.00**	1.00	0.05	1.00	0.36
QUAD	male2	QUAD	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD	male2	QUAD	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD	male2	QUAD	male1	0.81	**0.00**	0.92	**0.00**	0.97	**0.00**	0.99	**0.00**	1.00	**0.00**	1.00	0.16	1.00	0.86
QUAD x COLZ	female2	QUAD x COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	male1	QUAD x COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	male1	QUAD x COLZ	female2	1.00	0.99	1.00	0.98	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	male2	QUAD x COLZ	female1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	male2	QUAD x COLZ	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	male2	QUAD x COLZ	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 23 The results of a Dunn's post-hoc test for pair-wise comparisons within species or hybrids of the *An. coluzzii* - *An. quandriannulatus* species comparison. P-values are reported for X:A and 2:3 expression ratios for increasing RPKM cut-offs. Significant p-values (<0.05) are shown in bold font and indicate significant differences in the median of the distributions of the two samples in question (spp1 sex 1, spp2 sex 2).

				RPKM > 0.0		RPKM > 0.2		RPKM > 0.5		RPKM > 1.0		RPKM > 2.0		RPKM > 5.0		RPKM > 10.0	
spp1	sex1	spp2	sex2	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3	X/A	2/3
COLZ	female1	COLZ	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	female1	COLZ	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	female1	COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	female2	COLZ	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	female2	COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ	male1	COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	female1	COLZ x QUAD	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	female1	COLZ x QUAD	male1	1.00	1.00	0.79	1.00	0.44	1.00	0.29	1.00	0.99	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	female1	COLZ x QUAD	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	female2	COLZ x QUAD	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	female2	COLZ x QUAD	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
COLZ x QUAD	male1	COLZ x QUAD	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD	female1	QUAD	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD	female1	QUAD	male1	1.00	0.08	1.00	0.18	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD	female1	QUAD	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD	female2	QUAD	male1	0.69	**0.01**	1.00	0.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD	female2	QUAD	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD	male1	QUAD	male2	1.00	**0.00**	1.00	**0.01**	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	female1	QUAD x COLZ	female2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	female1	QUAD x COLZ	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	female1	QUAD x COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	female2	QUAD x COLZ	male1	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	female2	QUAD x COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
QUAD x COLZ	male1	QUAD x COLZ	male2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 24 Results of an ANOVA test (F-value and p-values, Pr>(F)) for differences between the mean of the X:A and 2:3 expression ratio distributions within each sample of the the *An. coluzzii* - *An. quadriannulatus* species comparison at increasing minimum RPKM cut-off thresholds. Significant p-values (≤0.05) are shown in bold font.

RPKM Cut-Off			> 0.0		> 0.2		> 0.5		> 1.0		> 2.0		> 5.0		> 10.0	
Species / F1 Hybrid	Sex	Biological Replicate	F-value	Pr(>F)	F-value	Pr(>F)	F-value	Pr(>F)	F-value	Pr(>F)	F-value	Pr(>F)	F-value	Pr(>F)	F-value	Pr(>F)
QUAD	Female	1	4.19	**0.041**	4.27	**0.039**	3.73	0.053	3.44	0.064	2.75	0.098	2.02	0.155	0.90	0.343
QUAD	Female	2	4.59	**0.032**	4.78	**0.029**	4.11	**0.043**	3.51	0.061	3.06	0.080	2.10	0.148	0.89	0.346
QUAD	Male	1	4.73	**0.030**	4.59	**0.032**	4.39	**0.036**	4.01	**0.045**	2.91	0.088	2.14	0.144	0.98	0.324
QUAD	Male	2	5.35	**0.021**	5.28	**0.022**	4.67	**0.031**	4.03	**0.045**	3.47	0.063	2.45	0.117	1.02	0.313
COLZ	Female	1	2.51	0.113	2.27	0.132	1.95	0.162	1.09	0.296	0.56	0.454	0.18	0.676	0.00	0.955
COLZ	Female	2	2.36	0.124	2.21	0.137	1.59	0.208	0.99	0.321	0.49	0.485	0.17	0.682	0.00	0.995
COLZ	Male	1	4.05	**0.044**	3.85	**0.050**	3.24	0.072	2.36	0.125	1.18	0.277	0.80	0.372	0.12	0.731
COLZ	Male	2	3.10	0.078	2.64	0.104	1.94	0.163	1.22	0.269	0.73	0.394	0.28	0.597	0.01	0.924
QUAD x COLZ	Female	1	3.58	0.059	3.37	0.067	2.85	0.092	2.09	0.148	1.23	0.267	0.56	0.456	0.12	0.726
QUAD x COLZ	Female	2	4.06	**0.044**	3.84	**0.050**	3.13	0.077	2.50	0.114	1.49	0.222	0.76	0.385	0.20	0.657
QUAD x COLZ	Male	1	3.38	0.066	3.29	0.070	2.58	0.108	1.79	0.181	1.10	0.295	0.44	0.509	0.07	0.786
QUAD x COLZ	Male	2	4.25	**0.039**	4.14	**0.042**	3.38	0.066	2.30	0.129	1.56	0.212	0.79	0.373	0.24	0.626
COLZ x QUAD	Female	1	1.43	0.232	1.26	0.261	0.82	0.364	0.41	0.525	0.38	0.537	0.05	0.821	0.03	0.858
COLZ x QUAD	Female	2	1.53	0.216	1.58	0.209	1.09	0.296	0.46	0.497	0.38	0.538	0.02	0.881	0.05	0.829
COLZ x QUAD	Male	1	2.19	0.139	2.09	0.148	1.63	0.202	1.13	0.288	0.65	0.419	0.21	0.649	0.01	0.934
COLZ x QUAD	Male	2	1.74	0.187	1.61	0.205	1.25	0.263	0.70	0.404	0.46	0.499	0.07	0.792	0.00	0.996

Table 25 Results of the Kruskal-Wallis test (X^2 and p-values) for differences between the median of the X:A and 2:3 expression ratio distributions within each sample of the the *An. coluzzii* - *An. quadriannulatus* species comparison at increasing minimum RPKM cut-off thresholds. Significant p-values (≤0.05) are shown in bold font.

RPKM Cut-Off			> 0.0		> 0.2		> 0.5		> 1.0		> 2.0		> 5.0		> 10.0	
Species / F1 Hybrid	Sex	Biological Replicate	X^2	p-value	X^2	p-value	X^2	p-value	X^2	p-value	X^2	p-value	X^2	p-value	X^2	p-value
QUAD	Female	1	23.90	**0.000**	32.50	**0.000**	25.50	**0.000**	18.76	**0.000**	13.60	**0.000**	6.18	**0.013**	0.00	0.983
QUAD	Female	2	21.61	**0.000**	30.81	**0.000**	23.17	**0.000**	14.82	**0.000**	11.70	**0.001**	4.48	**0.034**	0.23	0.629
QUAD	Male	1	48.16	**0.000**	56.57	**0.000**	57.54	**0.000**	50.33	**0.000**	31.92	**0.000**	15.75	**0.000**	2.15	0.143
QUAD	Male	2	48.41	**0.000**	56.29	**0.000**	44.75	**0.000**	31.13	**0.000**	26.52	**0.000**	11.62	**0.001**	0.03	0.857
COLZ	Female	1	16.76	**0.000**	19.02	**0.000**	17.39	**0.000**	8.09	**0.004**	1.39	0.238	0.37	0.542	0.67	0.411
COLZ	Female	2	11.60	**0.001**	14.10	**0.000**	9.60	**0.002**	4.40	**0.036**	0.21	0.646	0.01	0.915	1.13	0.287
COLZ	Male	1	42.02	**0.000**	46.19	**0.000**	40.53	**0.000**	27.34	**0.000**	9.10	**0.003**	5.49	**0.019**	0.18	0.673
COLZ	Male	2	34.35	**0.000**	32.90	**0.000**	25.19	**0.000**	15.31	**0.000**	6.83	**0.009**	2.89	0.089	0.05	0.829
QUAD x COLZ	Female	1	25.24	**0.000**	28.63	**0.000**	22.78	**0.000**	12.21	**0.000**	2.30	0.130	0.15	0.700	0.60	0.440
QUAD x COLZ	Female	2	24.82	**0.000**	28.79	**0.000**	21.65	**0.000**	13.26	**0.000**	2.76	0.097	0.23	0.628	0.34	0.559
QUAD x COLZ	Male	1	59.42	**0.000**	58.56	**0.000**	44.44	**0.000**	27.20	**0.000**	11.42	**0.001**	3.46	0.063	0.01	0.916
QUAD x COLZ	Male	2	62.89	**0.000**	63.99	**0.000**	50.09	**0.000**	28.15	**0.000**	13.35	**0.000**	4.08	**0.043**	0.04	0.836
COLZ x QUAD	Female	1	6.35	**0.012**	5.72	**0.017**	1.52	0.217	0.01	0.940	1.70	0.192	1.86	0.172	4.42	**0.036**
COLZ x QUAD	Female	2	8.51	**0.004**	10.69	**0.001**	4.64	**0.031**	0.17	0.681	0.68	0.410	1.86	0.172	3.77	0.052
COLZ x QUAD	Male	1	25.52	**0.000**	30.35	**0.000**	21.60	**0.000**	13.32	**0.000**	3.50	0.062	0.22	0.641	1.22	0.270
COLZ x QUAD	Male	2	20.64	**0.000**	22.84	**0.000**	16.11	**0.000**	6.96	**0.008**	1.44	0.229	0.03	0.861	1.17	0.279

Table 26 Comparison of M:F expression ratio distributions, separated by X-linked and autosomal genes. Genes are only included if they have an expression level greater than 10.0 RPKM in all samples within a comparison (*An. coluzzii* - *An. arabiensis* or *An. coluzzii* - *An. quadriannulatus*). F-values and p-values (Pr(<F)) are reported for an ANOVA, which was used to test for significant differences between the means of the X-linked and autosomal M:F expression ratio distributions. X^2 and p-values are reported for the Kruskal-Wallis test, which was used to test for significant differences between the medians of the X-linked and autosomal M:F expression ratio distributions.

	X-linked		Autosomal		ANOVA		Kruskal-Wallis Test	
Species / F1 Hybrid	Median	95% CI	Median	95% CI	F-value	Pr(<F)	X^2	p-value
COLZ	0.99	(0.97,1)	1.05	(1.05,1.06)	0.09	0.904	47.29	0.000
ARAB	0.91	(0.87,0.95)	0.91	(0.9,0.92)	0.05	0.997	6.79	0.005
COLZ x ARAB	0.95	(0.93,0.97)	0.98	(0.97,0.98)	0.04	0.999	7.45	0.003
ARAB x COLZ	0.90	(0.84,0.95)	1.00	(0.99,1.02)	0.07	0.968	6.59	0.005
COLZ	0.99	(0.97,1)	1.04	(1.04,1.05)	0.08	0.993	35.16	0.000
QUAD	0.89	(0.83,0.94)	0.96	(0.95,0.97)	0.05	0.999	25.54	0.000
COLZ x QUAD	1.07	(1.03,1.11)	1.11	(1.1,1.12)	0.10	0.976	10.09	0.001
QUAD x COLZ	0.87	(0.85,0.89)	0.90	(0.89,0.9)	0.05	0.999	13.00	0.000

Table 27 Number and percentage of male- and female-biased genes in *An. arabiensis* (ARAB), *An. coluzzii* (COLZ), and *An. quadriannulatus* (QUAD) per chromosome arm. The percentages provided are the percentage of genes on a chromosome arm that are sex-biased (ex, (num. sex-biased on 2L / total genes on 2L) * 100), rather than percentage of sex-biased genes out of the total gene set. P-values are provided for chi-square tests of the non-random distribution of mis-expressed genes among chromosome arms.

Chromosome Arm	Total	ARAB Male-biased		ARAB Female-biased		COLZ Male-biased		COLZ Female-biased		QUAD Male-biased		QUAD Female-biased	
2L	3,630	101	2.78%	10	0.28%	103	2.84%	21	0.58%	92	2.53%	29	0.80%
2R	4,598	116	2.52%	13	0.28%	124	2.70%	12	0.26%	129	2.81%	11	0.24%
3L	2,403	62	2.58%	12	0.50%	69	2.87%	11	0.46%	66	2.75%	21	0.87%
3R	3,118	140	4.49%	10	0.32%	121	3.88%	12	0.38%	112	3.59%	13	0.42%
X	1,309	13	0.99%	25	1.91%	15	1.15%	9	0.69%	17	1.30%	24	1.83%
Total	15,058	432	2.87%	70	0.46%	432	2.87%	65	0.43%	416	2.76%	98	0.65%
Chi-square	-	**1.30E-09**		**1.35E-13**		**4.68E-05**		1.31E-01		**9.27E-04**		**2.74E-09**	

Table 28 Results of a gene ontology (GO) term enrichment analysis for female-biased genes for each parental species. Note: *An. arabiensis* is not included because no significant results were found. GO types include biological process (BP), cellular component (CC), molecular function (MF), and KEGG functional pathways (keg).

Species Female	Num. Searched	Num. GO Genes	Obs. GO Genes	Corrected p-value	GO Term ID	GO Type	GO Term
COLZ	40	136	7	0.001	GO:1901565	BP	organonitrogen compound catabolic process
COLZ	40	31	7	0.000	GO:0043171	BP	peptide catabolic process
COLZ	40	693	14	0.000	GO:0071944	CC	cell periphery
COLZ	40	676	14	0.000	GO:0005886	CC	plasma membrane
COLZ	40	55	6	0.000	GO:0033218	MF	amide binding
COLZ	40	50	6	0.000	GO:0042277	MF	peptide binding
COLZ	40	106	7	0.000	GO:0008238	MF	exopeptidase activity
COLZ	40	50	6	0.000	GO:0004177	MF	aminopeptidase activity
COLZ	40	145	7	0.001	GO:0008237	MF	metallopeptidase activity
COLZ	40	69	7	0.000	GO:0008235	MF	metalloexopeptidase activity
COLZ	40	28	6	0.000	GO:0070006	MF	metalloaminopeptidase activity
COLZ	7	12	2	0.008	KEGG:00730	keg	Thiamine metabolism
COLZ	7	16	2	0.014	KEGG:00790	keg	Folate biosynthesis
QUAD	12	16	2	0.039	KEGG:00790	keg	Folate biosynthesis

Table 29 Results of a gene ontology (GO) term enrichment analysis for male-biased genes for each parental species. GO types include biological process (BP), cellular component (CC), molecular function (MF), KEGG functional pathways (keg), and miRBase microRNAs (mi). Table continues on subsequent pages.

Species Male	Num. Searched	Num. GO Genes	Obs. GO Genes	Corrected p-value	GO Term ID	GO Type	GO Term
ARAB	226	6	5	0.000	GO:0001578	BP	microtubule bundle formation
ARAB	226	61	9	0.007	GO:0030030	BP	cell projection organization
ARAB	226	11	6	0.000	GO:0030031	BP	cell projection assembly
ARAB	226	23	8	0.000	GO:0044782	BP	cilium organization
ARAB	226	39	11	0.000	GO:0060271	BP	cilium assembly
ARAB	226	6	5	0.000	GO:0035082	BP	axoneme assembly
ARAB	226	4	3	0.036	GO:0070286	BP	axonemal dynein complex assembly
ARAB	226	102	18	0.000	GO:0006928	BP	movement of cell or subcellular component
ARAB	226	131	25	0.000	GO:0007017	BP	microtubule-based process
ARAB	226	58	16	0.000	GO:0007018	BP	microtubule-based movement
ARAB	226	5	4	0.001	GO:0003341	BP	cilium movement
ARAB	226	209	25	0.000	GO:0005856	CC	cytoskeleton
ARAB	226	128	22	0.000	GO:0015630	CC	microtubule cytoskeleton
ARAB	226	56	13	0.000	GO:0042995	CC	cell projection
ARAB	226	39	13	0.000	GO:0005929	CC	cilium
ARAB	226	25	10	0.000	GO:0044463	CC	cell projection part
ARAB	226	170	25	0.000	GO:0044430	CC	cytoskeletal part
ARAB	226	55	13	0.000	GO:0005875	CC	microtubule associated complex
ARAB	226	22	10	0.000	GO:0044441	CC	ciliary part
ARAB	226	13	8	0.000	GO:0097014	CC	ciliary plasm
ARAB	226	13	8	0.000	GO:0005930	CC	axoneme
ARAB	226	5	4	0.001	GO:0044447	CC	axoneme part
ARAB	226	22	13	0.000	GO:0030286	CC	dynein complex
ARAB	226	4	4	0.000	GO:0005858	CC	axonemal dynein complex
ARAB	226	2370	81	0.023	GO:0005515	MF	protein binding
ARAB	226	59	11	0.000	GO:0003774	MF	motor activity
ARAB	226	37	11	0.000	GO:0003777	MF	microtubule motor activity
ARAB	16	31	3	0.013	KEGG:00330	keg	Arginine and proline metabolism
ARAB	250	241	17	0.012	MI:dme-miR-	mi	MI:dme-miR-1011

Table 29, continued.

Species Male	Num. Searched	Num. GO Genes	Obs. GO Genes	Corrected p-value	GO Term ID	GO Type	GO Term
COLZ	237	102	16	0.000	GO:0006928	BP	movement of cell or subcellular component
COLZ	237	131	24	0.000	GO:0007017	BP	microtubule-based process
COLZ	237	58	15	0.000	GO:0007018	BP	microtubule-based movement
COLZ	237	5	4	0.001	GO:0003341	BP	cilium movement
COLZ	237	6	5	0.000	GO:0001578	BP	microtubule bundle formation
COLZ	237	11	6	0.000	GO:0030031	BP	cell projection assembly
COLZ	237	23	7	0.001	GO:0044782	BP	cilium organization
COLZ	237	39	12	0.000	GO:0060271	BP	cilium assembly
COLZ	237	6	5	0.000	GO:0035082	BP	axoneme assembly
COLZ	237	4	3	0.044	GO:0070286	BP	axonemal dynein complex assembly
COLZ	237	209	26	0.000	GO:0005856	CC	cytoskeleton
COLZ	237	128	23	0.000	GO:0015630	CC	microtubule cytoskeleton
COLZ	237	56	14	0.000	GO:0042995	CC	cell projection
COLZ	237	39	14	0.000	GO:0005929	CC	cilium
COLZ	237	25	11	0.000	GO:0044463	CC	cell projection part
COLZ	237	170	26	0.000	GO:0044430	CC	cytoskeletal part
COLZ	237	55	14	0.000	GO:0005875	CC	microtubule associated complex
COLZ	237	22	11	0.000	GO:0044441	CC	ciliary part
COLZ	237	13	7	0.000	GO:0097014	CC	ciliary plasm
COLZ	237	13	7	0.000	GO:0005930	CC	axoneme
COLZ	237	5	4	0.001	GO:0044447	CC	axoneme part
COLZ	237	22	13	0.000	GO:0030286	CC	dynein complex
COLZ	237	4	4	0.000	GO:0005858	CC	axonemal dynein complex
COLZ	237	59	11	0.000	GO:0003774	MF	motor activity
COLZ	237	37	11	0.000	GO:0003777	MF	microtubule motor activity
COLZ	237	2370	84	0.028	GO:0005515	MF	protein binding
COLZ	19	31	3	0.025	KEGG:00330	keg	Arginine and proline metabolism

Table 29, continued.

Species Male	Num. Searched	Num. GO Genes	Obs. GO Genes	Corrected p-value	GO Term ID	GO Type	GO Term
QUAD	189	5	3	0.050	GO:0001539	BP	cilium or flagellum-dependent cell motility
QUAD	189	102	16	0.000	GO:0006928	BP	movement of cell or subcellular component
QUAD	189	131	24	0.000	GO:0007017	BP	microtubule-based process
QUAD	189	58	15	0.000	GO:0007018	BP	microtubule-based movement
QUAD	189	5	5	0.000	GO:0003341	BP	cilium movement
QUAD	189	61	8	0.014	GO:0030030	BP	cell projection organization
QUAD	189	11	7	0.000	GO:0030031	BP	cell projection assembly
QUAD	189	23	7	0.000	GO:0044782	BP	cilium organization
QUAD	189	39	11	0.000	GO:0060271	BP	cilium assembly
QUAD	189	63	8	0.018	GO:0000226	BP	microtubule cytoskeleton organization
QUAD	189	6	6	0.000	GO:0001578	BP	microtubule bundle formation
QUAD	189	6	6	0.000	GO:0035082	BP	axoneme assembly
QUAD	189	4	4	0.000	GO:0070286	BP	axonemal dynein complex assembly
QUAD	189	3	3	0.005	GO:0036158	BP	outer dynein arm assembly
QUAD	189	209	24	0.000	GO:0005856	CC	cytoskeleton
QUAD	189	128	21	0.000	GO:0015630	CC	microtubule cytoskeleton
QUAD	189	56	13	0.000	GO:0042995	CC	cell projection
QUAD	189	39	12	0.000	GO:0005929	CC	cilium
QUAD	189	25	10	0.000	GO:0044463	CC	cell projection part
QUAD	189	170	24	0.000	GO:0044430	CC	cytoskeletal part
QUAD	189	55	13	0.000	GO:0005875	CC	microtubule associated complex
QUAD	189	22	10	0.000	GO:0044441	CC	ciliary part
QUAD	189	13	7	0.000	GO:0097014	CC	ciliary plasm
QUAD	189	13	7	0.000	GO:0005930	CC	axoneme
QUAD	189	5	4	0.000	GO:0044447	CC	axoneme part
QUAD	189	22	13	0.000	GO:0030286	CC	dynein complex
QUAD	189	4	4	0.000	GO:0005858	CC	axonemal dynein complex
QUAD	189	2370	77	0.000	GO:0005515	MF	protein binding
QUAD	189	59	11	0.000	GO:0003774	MF	motor activity
QUAD	189	37	11	0.000	GO:0003777	MF	microtubule motor activity
QUAD	18	31	3	0.018	KEGG:00330	keg	Arginine and proline metabolism

QUAD	214	241	17	0.002	MI:dme-miR-1011	mi	MI:dme-miR-1011
QUAD	214	367	19	0.034	MI:aga-miR-124	mi	MI:aga-miR-124

Table 30 Number of genes mis-expressed in F1 hybrid females per chromosome arm. Each chromosome arm is further divided into genes that show patterns of additive, over-, or under-expression in hybrids compared to parental strains of the same sex. Additionally, the number of mis-expressed genes that are sex-biased in parental species is reported.

		ARAB x COLZ Female		COLZ x ARAB Female		QUAD x COLZ Female		COLZ x QUAD Female	
Chrom Arm	Pattern	All Mis-Expressed	Sex-Biased	All Mis-Expressed	Sex-Biased	All Mis-Expressed	Sex-Biased	All Mis-Expressed	Sex-Biased
2L	Additive	24	1	20	2	45	3	31	4
	Over	45	1	28	1	55	6	114	9
	Under	404	7	367	8	214	2	278	6
2R	Additive	16	1	12	0	24	0	15	0
	Over	75	1	50	1	54	4	154	7
	Under	502	9	470	12	288	2	383	10
3L	Additive	20	2	9	1	21	2	6	1
	Over	14	0	8	1	15	1	53	7
	Under	120	4	113	3	57	3	136	1
3R	Additive	7	1	9	0	20	1	5	0
	Over	35	1	26	0	35	3	110	7
	Under	322	5	280	4	157	2	225	3
X	Additive	3	0	4	1	7	2	8	1
	Over	24	1	17	1	15	4	47	5
	Under	159	4	147	4	93	0	99	2
Total		1,770	38	1,560	39	1,100	35	1,664	63
Percentage of Total Genes		11.75%	0.25%	10.36%	0.26%	7.31%	0.23%	11.05%	0.42%

Table 31 Number of genes mis-expressed in F1 hybrid males per chromosome arm. Each chromosome arm is further divided into genes that show patterns of additive, over-, or under-expression in hybrids compared to parental strains of the same sex. Additionally, the number of mis-expressed genes that are sex-biased in parental species is reported.

		ARAB x COLZ Male		COLZ x ARAB Male		QUAD x COLZ Male		COLZ x QUAD Male	
Chrom Arm	Pattern	All Mis-Expressed	Sex-Biased	All Mis-Expressed	Sex-Biased	All Mis-Expressed	Sex-Biased	All Mis-Expressed	Sex-Biased
2L	Additive	21	1	14	0	32	2	17	1
	Over	125	4	26	1	73	9	139	14
	Under	302	14	369	63	219	3	287	24
2R	Additive	10	3	9	2	20	3	2	0
	Over	202	3	54	0	77	3	165	6
	Under	422	26	508	76	296	4	373	24
3L	Additive	16	3	13	4	13	4	13	2
	Over	69	3	5	1	24	2	61	5
	Under	108	13	126	36	71	2	135	16
3R	Additive	13	6	7	2	19	4	10	1
	Over	82	2	25	0	82	15	153	17
	Under	256	17	323	79	152	1	217	19
X	Additive	2	0	2	0	9	3	3	0
	Over	54	1	17	4	22	4	50	3
	Under	148	3	139	3	112	2	96	0
Total		1,830	99	1,637	271	1,221	61	1,721	132
Percentage of Total Genes		12.15%	0.66%	10.87%	1.80%	8.11%	0.41%	11.43%	0.88%

Table 32 Percentage of mis-expressed genes in F1 hybrids that show a pattern of additive, over-, or under- expression.

	ARAB x COLZ		COLZ x ARAB		QUAD x COLZ		COLZ x QUAD	
	Male	Female	Male	Female	Male	Female	Male	Female
Additive	3.39%	3.95%	2.75%	3.46%	7.62%	10.64%	2.61%	3.91%
Over	29.07%	10.90%	7.76%	8.27%	22.77%	15.82%	33.00%	28.73%
Under	67.54%	85.14%	89.49%	88.27%	69.62%	73.55%	64.38%	67.37%
Total mis-expressed genes	1,830	1,770	1,637	1,560	1,221	1,100	1,721	1,664

Table 33 The results of a gene ontology search for genes internal to the interaction network of mis-expressed genes in ARAB x COLZ female F1 hybrids.

ARAB x COLZ Female *An. gambiae* **Gene**	**Drosophila Ortholog**	**Function Summary**
	CG30281	Predicted chitin-binding properties
AGAP007041	CG5550	Fibrinogen-like encoding gene that is up-regulated after injury in *D. melanogaster*, involved in cell adhesion and may have clotting properties.
	CG6788	Fibrinogen-like encoding gene.
AGAP007280	NDL	Nudel is a multi-functional protein with serine protease activity, required for eggshell biogenesis and embryonic dorso-ventral patterning.
AGAP011305	CG1809	Alkaline phosphatase
AGAP028116	CG6617	No data.
AGAP028124	CG13138	Putative *An. gambiae* cuticular protein CPLCP2. *D. melanogaster* homolog expressed downstream of DSX in females.

Table 34 The results of a gene ontology search for genes internal to the interaction network of mis-expressed genes in COLZ x ARAB female F1 hybrids.

COLZ x ARAB Female *An. gambiae* **Gene**	**Drosophila Ortholog**	**Function Summary**
AGAP000219	CG17450	TEKT3 is a member of the tektin family of microtubule-associated cytoskeletal proteins primarily expressed in male germ cells, may be involved in sperm competition. Tektins are insoluble, filament-forming proteins essential for the construction of cilia and flagella. They can be found in ciliary and flagellar axonemes, basal bodies and centrioles and may play a role in flagellum stability and sperm motility.
	CG6788	Fibrinogen-like encoding gene.
AGAP007041	CG5550	Fibrinogen-like encoding gene that is up regulated after injury in *D. melanogaster*, involved in cell adhesion and may have clotting properties.
	CG30281	Predicted chitin-binding properties
AGAP007280	NDL	Nudel is a multi-functional protein with serine protease activity, required for eggshell biogenesis and embryonic dorsoventral patterning.
AGAP011750	CG13397	Alpha-N-acetylglucosaminidase (NAGLU) is a lysosomal enzyme required for the stepwise degradation of heparan sulphate. NAGLU mutations can lead to neurological dysfunction.

Table 35 The results of a gene ontology search for genes internal to the interaction network of mis-expressed genes in QUAD x COLZ female F1 hybrids.

QUAD x COLZ Female *An. gambiae* Gene	Drosophila Ortholog	Function Summary
AGAP004787	FarO	Fatty acyl-CoA reductase in oenocytes (FarO) (along with Kar) regulates the balance between cell growth and lipid storage of larval oenocytes. In *D. melanogaster*, larval oenocytes synthesize fatty acids required for tracheal waterproofing, and adult oenocytes produce cuticular hyrocarbons required for desiccation resistance and pheromonal communication.
AGAP005435	IYD	No data.
AGAP006891	ND-49	NADH dehydrogenase (ubiquinone) 49 kDa subunit (ND-49) is one of the "entry enzymes" of cellular respiration or oxidative phosphorylation in the mitochondria.
AGAP006968	CG8177	Anion exchange protein / transporter, regulates intracellular pH. CG8177 knockdown significantly increased the frequency of multilayered follicular stalks in *D. melanogaster*.
AGAP007280	NDL	Nudel is a multi-functional protein with serine protease activity, required for eggshell biogenesis and embryonic dorsoventral patterning.
AGAP009871	CPR49AE	Inferred chitin-based cuticle development.
AGAP011220	AUST	Australin (Aust) is a male meiotic specific paralogue of borr, which is one of the three targeting subunits for aurB kinase in the chromosomal passenger complex (CPC). The CPC is critical in regulating multiple aspects of cell division, including chromosome condensation, kinetochore function and cytokinesis, through the kinase activity of aurB.

	BORR	Borealin-related is one of the three targeting subunits for aurB kinase in the Chromosomal passenger complex. It helps to target the complex to the centromere region of chromosomes and the cleavage furrow during cytokinesis. At centromeres the complex is involved both in pausing mitotic progression when there are chromosome mis-attachments and in correcting those mis-attachments. At the cleavage furrow, the complex is involved in regulating the process of abscission (cell separation).
AGAP011305	CG1809	Alkaline phosphatase.
AGAP028563	CG6785	Involved in heat stress response, possible heat stress protein, regulated by the transcription factor Heat Shock Factor".

Table 36 The results of a gene ontology search for genes internal to the interaction network of mis-expressed genes in COLZ x QUAD female F1 hybrids.

COLZ x QUAD Female *An. gambiae* **Gene**	**Drosophila Ortholog**	**Function Summary**
AGAP000751	MAD2	Mad2 is a conserved component of the spindle checkpoint. During mitosis, it is recruited to unattached kinetochores, where it binds Mad1 and fzy, promoting the assembly of the mitotic checkpoint complex. During interphase, Mad2 is intranuclear, bound to Mad1, and associated primarily with the nuclear pore complex.
AGAP002075	ALD/Mps1	Monopolar spindle 1 (Mps1) is necessary for the proper functioning of the mitotic and meiotic spindle checkpoints (MSCs), which monitor the integrity of the spindle apparatus and prevent cells from progressing into anaphase until chromosomes are properly aligned on the metaphase plate
AGAP003227	NMDYN-D7	Nucleoside diphosphate kinases (NDK) are enzymes required for the synthesis of nucleoside triphosphates (NTP) other than ATP. They provide NTPs for nucleic acid synthesis, CTP for lipid synthesis, UTP for polysaccharide synthesis and GTP for protein elongation, signal transduction and microtubule polymerization.
AGAP006387	DNAPOL-ALPHA60	DNA primase is the polymerase that synthesizes small RNA primers for the Okazaki fragments made during discontinuous DNA replication.
AGAP006891	ND-49	NADH dehydrogenase (ubiquinone) 49 kDa subunit (ND-49) is one of the "entry enzymes" of cellular respiration or oxidative phosphorylation in the mitochondria.
AGAP007779	SMYDA-3	Smyd proteins are important in epigenetic control of development and carcinogenesis through posttranslational modifications in histones and other proteins.
AGAP009871	CPR49AE	Chitin-based cuticle development.
AGAP010901	CG7548	Cuticular protein.
AGAP010901	CG8541	Cuticular protein.

Table 37 The results of a gene ontology search for genes internal to the interaction network of mis-expressed genes in ARAB x COLZ male F1 hybrids.

ARAB x COLZ Male *An. gambiae* **Gene**	**Drosophila Ortholog**	**Function Summary**
AGAP000416	CG10252	Human ortholog: ODF3 aka SHIPPO1, outer dense fiber of sperm tails 3.
AGAP001112	CG11449	Cilia and flagella associated protein 45.
AGAP001620	CG11893	Involved in wing disk development.
AGAP002033	CG16721	Arganine kinase. Human ortholog: T-complex 11, testis-specific-like 1 (TCP11L1) is only expressed in fertile adult mammalian testes and is thought to be important in sperm function and fertility.
AGAP003083	CG17018	Lipase 3. Human ortholog: Meiosis arrest female 1 (MARF1), essential for oogenesis in mice. Mutations of Marf1 cause female infertility characterized by up-regulation of a cohort of transcripts, increased retrotransposon expression, defective cytoplasmic maturation, and meiotic arrest.
	CG17097	Drosophila accessory gland / seminal fluid protein.
	CG18284	Drosophila accessory gland / seminal fluid protein.
AGAP005309	CG18301	No data.
AGAP005908	CG3213	Human ortholog: ODF2, outer dense fiber of sperm tails 2.
AGAP007364	CG4365	Drosophila ortholog: Tenzing norgay (TZN) , involved in hypoxia tolerance.
AGAP007823	CG4476	Drosophila ortholog: CG4476, an orphan member of the SLC6 family of transporters that includes the plasma membrane serotonin and dopamine transporters. CG4476 mutants show a decreased behavioral response to light.
AGAP008186	CG4546	Drosophila accessory gland / seminal fluid protein.
AGAP009151	CG7387	Human ortholog: DnaJ, heat shock protein family (Hsp40) member A3
AGAP010860	EXU	Drosophila ortholog: Exuperantia (EXU), required for Drosophila spermatogenesis as well as anteroposterior polarity of the developing oocyte, and encodes overlapping sex-specific transcripts. Tra-2 is required in male germ cells for efficient male-specific processing of exu RNA; in the absence of tra-2, X/Y males produce a new exu mRNA which is processed at its 3' end so that it contains sequences normally specific to the female 3' untranslated region.

AGAP013346	ELBA2	Drosophila ortholog: early boundary activity 2 (ELBA2) is involved in positive regulation of chromatin silencing.

Table 38 The results of a gene ontology search for genes internal to the interaction network of mis-expressed genes in COLZ x ARAB male F1 hybrids.

COLZ x ARAB Male *An. gambiae* **Gene**	**Drosophila Ortholog**	**Function Summary**
AGAP001620	CG10252	Human ortholog: ODF3 aka SHIPPO1, outer dense fiber of sperm tails 3.
AGAP007334	TEKTIN-C	Tektins are insoluble, filament-forming proteins essential for the construction of cilia and flagella. They can be found in ciliary and flagellar axonemes, basal bodies and centrioles and may play a role in flagellum stability and sperm motility.
AGAP009088	LIM3	Lim3 is a homeobox transcription factor that regulates neuronal sub-type identity, including motor neuron identity.
AGAP008341	LOK	*Drosophila* ortholog: Loki. Human ortholog: Checkpoint kinase 2 (chek2/chk2). CHK2 kinase is key component of double-stranded break repair during meiosis.
AGAP001112	CG11449	Cilia and flagella associated protein 45.
AGAP009292	CG12857	No data.
AGAP028124	CG13138	Putative An. gambiae cuticular protein CPLCP2. *D. melanogaster* ortholog expressed downstream of DSX in females.
AGAP007664	CG14011	*An. gambiae* coiled-coil domain-containing protein 3. *Drosophila* ortholog involved in cell-cell signaling involved in cell fate commitment, lateral inhibition.
AGAP011743	CG14183	Human ortholog: DRC-11, involved in axoneme assembly. The nexin–dynein regulatory complex (N-DRC) is proposed to coordinate dynein arm activity and interconnect doublet microtubules.
AGAP005805	CG15580	No data.
AGAP007823	CG17018	Lipase 3. Human orthlog: Meiosis arrest female 1 (MARF1), essential for oogenesis in mice. Mutations of Marf1 cause female infertility characterized by up-regulation of a cohort of transcripts, increased retrotransposon expression, defective cytoplasmic maturation, and meiotic arrest.
AGAP028457	BIP2/TAF3	*Drosophila* ortholog: TAF3 is a transcription factor involved in DNA damage response. It is a negative regulator of p53 transcription activation function. p53 induces DNA repair or programmed cell death by activating expression of its target genes. p53 is phosphorylated by Loki/CHK2, which is necessary for p53 mediated apoptosis.

	PPI1	Protein phosphatase 1c interacting protein 1
AGAP013414	CG31029	No data.
	CG3199	No data.
AGAP008186	CG3213	Human ortholog: ODF2, outer dense fiber of sperm tails 2
AGAP005880	CG32392	Human ortholog: radial spoke 3 ortholog (RSPH3). Radial spokes regulate the activity of inner arm dyneinin cilia/flagella through protein phosphorylation and dephosphorylation.
AGAP006170	CG3528	No data.
AGAP001767	UQCR-C1	*Drosophila* ortholog: Ubiquinol-cytochrome c reductase core protein 1 (UQCR-C1), involved in cytoplasmic microtubule organization.
AGAP010031	TEKTIN-A	Tektins are insoluble, filament-forming proteins essential for the construction of cilia and flagella. They can be found in ciliary and flagellar axonemes, basal bodies and centrioles and may play a role in flagellum stability and sperm motility.
AGAP001735	CG7264	Human orhtolog: RIB43A domain with coiled-coils 2. Involved in organization of ciliary and flagellar microtubules.
AGAP005908	CG7387	*Drosophila* ortholog: DnaJ heat shock protein family (Hsp40) member A3
AGAP001353	CG8997	*Drosophila* ortholog is member of the haemolymph juvenile hormone binding protein family.
AGAP007678	FEST	*Drosophila* ortholog: FEST. During spermatogenesis Rbp4 and FEST function to direct cell type- and stage-specific repression of translation of the core G2/M cell cycle component cycB during the specialized cell cycle of male meiosis.

Table 39 The results of a gene ontology search for genes internal to the interaction network of mis-expressed genes in QUAD x COLZ male F1 hybrids.

QUAD x COLZ Male *An. gambiae* **Gene**	**Drosophila Ortholog**	**Function Summary**
AGAP005756	CLT	*Drosophila* ortholog: cricklet (CLT). CLT mutants are defective in yolk protein synthesis, histolysis of the larval fat body, vitellogenesis, and synthesis of larval serum protein 2 in the adult. The *clt* locus may encode a protein essential for mediating the response of adult tissues to juvenile hormone.
AGAP011379	FZ	*Drosophila* ortholog: Frizzled (FZ), a member of the *WNT* signalling pathway, involved in planar cell polarity, which is the coordination of the cytoskeleton of epidermal cells to produce a parallel array of cuticular hairs and bristles
AGAP000835	OTU	*Drosophila* ortholog: Ovarian Tumor (OTU). Thiol-dependent ubiquitin-specific protease activity, required in oogenesis. Mutants develop tumerous ovarian follicles; germ cells fail to differentiate into nurse cells or oocytes.
AGAP012986	SPI	*Drosophila* ortholog: Spitz (Spi) is the cardinal Egfr ligand that is produced as a transmembrane precursor and processed by S and rho. Spi roles include growth regulation, cell survival and developmental patterning. Important during oogenesis, embryo development, wing morphogenesis.
AGAP004645	SCP2	*Drosophila* ortholog: Sarcoplasmic calcium-binding protein 2 (SCP2), involved in Ca ion regulation in the cytoplasm.
AGAP009871	CPR49AE	Cuticular protein 49Ae (CPR49AE) involved in chitin-based cuticle development.
AGAP001952	S-LAP1	*Drosophila* ortholog: Sperm-Leucylaminopeptidase 1 (S-LAP1), s-LAPS are specifically expressed in the testis and all encode proteins incorporated in mature sperm.
AGAP006968	CG8177	Anion exchange protein / transporter, regulates intracellular pH. CG8177 knockdown significantly increased the frequency of multilayered follicular stalks in *D. melanogaster*.
AGAP011512	CG7365	Implicated in wing morphogenesis in *D. melanogaster*.

AGAP012986	KRN	*Drosophila* ortholog: Keren (KRN) is an Egfr ligand that is processed by S and rho. It shows a redundant role with Spi in very few tissues. Important during oogenesis, embryo development, wing morphogenesis.
AGAP001952	S-LAP2	*Drosophila* ortholog: Sperm-Leucylaminopeptidase 2 (S-LAP2), s-LAPS are specifically expressed in the testis and all encode proteins incorporated in mature sperm.
AGAP007945	PPK13	*Drosophila* ortholog: Pickpocket 12 (PPK13). Pickpocket genes in Drosophila encode subunits of non-voltage gated, amiloride-sensitive cation channels. Channels may be formed by homo- or heteromeric arrangements of subunits. Each subunit has two transmembrane domains and a large cysteine-rich extracellular loop domain. They are functionally diverse, with roles in fluid and salt absorbance, mechanosensation and chemosensation.
AGAP013145	CPO	*Drosophila* ortholog: Couch potato (CPO), is most strongly expressed in nuclei of the central and peripheral nervous systems and the ring gland, where it may regulate gene transcription to control complex neurological and neuroendocrine functions. It is required for normal synaptic transmission, climate adaptation, olfaction, diapause and behavioral responses.

Table 40 The results of a gene ontology search for genes internal to the interaction network of mis-expressed genes in COLZ x QUAD male F1 hybrids.

COLZ x QUAD Male *An. gambiae* **Gene**	**Drosophila Ortholog**	**Function Summary**
AGAP001112	CG11449	Human ortholog: Cilia- and flagella-associated protein 45 (CFAP45 aka CCDC19, NESG1). May be involved in cell proliferation, apoptosis, and tumor suppression.
AGAP002075	ALD / MPS1	Monopolar spindle 1 (Mps1) is necessary for the proper functioning of the mitotic and meiotic spindle checkpoints (MSCs), which monitor the integrity of the spindle apparatus and prevent cells from progressing into anaphase until chromosomes are properly aligned on the metaphase plate
AGAP003227	NMDYN-D7	Nucleoside diphosphate kinases (NDK) are enzymes required for the synthesis of nucleoside triphosphates (NTP) other than ATP. They provide NTPs for nucleic acid synthesis, CTP for lipid synthesis, UTP for polysaccharide synthesis and GTP for protein elongation, signal transduction and microtubule polymerisation.
AGAP005098	SUN	Drosophila ortholog: Stunded (SUN), is a circulating insulinotropic peptide produced by fat cells. It modulates physiological insulin levels in response to nutrients and is required for normal spindle orientation during embryonic divisions
AGAP005805	CG15580	No data.
AGAP006098	NOS	*Drosophila* ortholog: Nanos (Nos) is an RNA-binding protein that forms part of a translational repressor complex. It functions as the localized determinant of abdominal segmentation. Nos contributes to germline development, germline stem cell renewal, and neuronal morphogenesis and function. In Drosophila, Nanos is maternally supplied and is involved in the migration of primordial germ cells to the gonad. In mice, Nanos2 is predominantly expressed in male germ cells, and the elimination of this gene results in a complete loss of spermatogonia.
AGAP006170	CG3528	No data.
AGAP007240	CG16984	Enkurin. In mammals, Enkurin is expressed at high levels in the testis, sperm, and vomeronasal organ. Enkurin interacts with TRPC cation channel proteins to facilitate Ca cation flux into sperm, a necessary step during fertilization.

AGAP007678	FEST	*Drosophila* ortholog: FEST. During spermatogenesis Rbp4 and FEST function to direct cell type- and stage-specific repression of translation of the core G2/M cell cycle component cycB during the specialized cell cycle of male meiosis.
AGAP008186	CG3213	Human ortholog: ODF2, outer dense fiber of sperm tails 2.
AGAP009026	CG7886	*Drosophila* ortholog: CG7886 contains a domain characteristic of centrosome proteins. The centrosome is the major microtubule-organizing centre of animal cells and through its influence on the cytoskeleton is involved in cell shape, polarity and motility. It also has a crucial function in cell division because it determines the poles of the mitotic spindle that segregates duplicated chromosomes between dividing cells.
AGAP009639	MIL	*Drosophila* ortholog: Milka/Hanabi. Hanabi is a member of the NAP familiy proteins, and is localized to the sperm cytoplasm. *Hanabi* mutants show scattered nuclei and abnormalities in nuclear shaping and spermatid elongation.
AGAP009871	CPR49AE	Cuticular protein 49Ae (CPR49AE) involved in chitin-based cuticle development.
AGAP010031	TEKTIN-A	Tektins are insoluble, filament-forming proteins essential for the construction of cilia and flagella. They can be found in ciliary and flagellar axonemes, basal bodies and centrioles and may play a role in flagellum stability and sperm motility.
AGAP010901	CG8541 CG7548	Putative cuticular protein. Putative cuticular protein.
AGAP011379	FZ	*Drosophila* ortholog: Frizzled (FZ), a member of the *WNT* signaling pathway, involved in planar cell polarity, which is the coordination of the cytoskeleton of epidermal cells to produce a parallel array of cuticular hairs and bristles

Table 41 This table reports the mode of allelic imbalance identified (if any) for autosomal genes in female hybrids. The intersection represents genes shared by both directions of the cross for each mode. P-values are reported for chi-square tests used to test for significant differences between the number of genes in each transcription category between directions of each cross.

Category	ARAB x COLZ Female		Intersection	COLZ x ARAB Female		p-value	QUAD x COLZ Female		Intersection	COLZ x QUAD Female		p-value
Conserved	8,971	72.07%	8,577	8,991	72.23%	0.88	4,258	34.21%	3,812	4,122	33.11%	0.14
Compensatory	1,358	10.91%	944	1,338	10.75%	0.70	1,092	8.77%	782	1,228	9.87%	0.00
Trans Only	1,525	12.25%	1,391	1,532	12.31%	0.90	5,758	46.26%	5,259	5,632	45.24%	0.24
Cis x Trans	331	2.66%	230	326	2.62%	0.85	870	6.99%	573	946	7.60%	0.07
Cis + Trans	214	1.72%	121	210	1.69%	0.85	400	3.21%	210	426	3.42%	0.37
Cis Only	49	0.39%	12	51	0.41%	0.84	70	0.56%	11	94	0.76%	0.06

Table 42 This table reports the mode of allelic imbalance identified (if any) for autosomal genes in male hybrids. The intersection represents genes shared by both directions of the cross for each mode. P-values are reported for chi-square tests used to test for significant differences between the number of genes in each transcription category between directions of each cross.

Category	ARAB x COLZ Male		Intersect	COLZ x ARAB Male		p-value	QUAD x COLZ Male		Intersect	COLZ x QUAD Male		p-value
Conserved	9,325	74.91%	8,812	9,498	76.30%	0.21	7,396	59.42%	6,853	7,388	59.35%	0.95
Compensatory	1,794	14.41%	1,108	1,621	13.02%	0.00	1,637	13.15%	1,102	1,645	13.21%	0.89
Trans Only	827	6.64%	684	832	6.68%	0.90	2,496	20.05%	2,213	2,454	19.71%	0.55
Cis x Trans	260	2.09%	150	250	2.01%	0.66	546	4.39%	376	603	4.84%	0.09
Cis + Trans	204	1.64%	94	197	1.58%	0.73	326	2.62%	163	299	2.40%	0.28
Cis Only	38	0.31%	5	50	0.40%	0.20	47	0.38%	6	59	0.47%	0.24

Table 43 This table reports the mode of allelic imbalance identified (if any) for X chromosome genes in female hybrids. The intersection represents genes shared by both directions of the cross for each mode. P-values are reported for chi-square tests used to test for significant differences between the number of genes in each transcription category between directions of each cross.

Category	ARAB x COLZ Female		Intersection	COLZ x ARAB Female		p-value	QUAD x COLZ Female		Intersection	COLZ x QUAD Female		p-value
Conserved	906	74.26%	869	894	73.28%	0.78	447	36.64%	412	434	35.57%	0.66
Compensatory	75	6.15%	50	87	7.13%	0.35	72	5.90%	50	85	6.97%	0.30
Trans Only	175	14.34%	156	164	13.44%	0.55	587	48.11%	554	577	47.30%	0.77
Cis x Trans	27	2.21%	22	31	2.54%	0.60	52	4.26%	35	58	4.75%	0.57
Cis + Trans	28	2.30%	14	33	2.70%	0.52	57	4.67%	32	54	4.43%	0.78
Cis Only	9	0.74%	0	11	0.90%	0.65	5	0.41%	0	12	0.98%	0.09

Table 44 This table reports the p-values for chi-square tests used to test for significant differences between the number of genes in each transcription category between female autosomal genes and X chromosome genes for each cross. Genes in each category were first corrected by the total number of genes on the autosomes or X chromosomes, respectively.

Category	ARAB x COLZ	COLZ x ARAB	QUAD x COLZ	COLZ x QUAD
Conserved	0.99	0.99	0.97	0.97
Compensatory	0.91	0.93	0.94	0.94
Trans Only	0.97	0.98	0.97	0.97
Cis x Trans	0.98	0.99	0.94	0.94
Cis + Trans	0.98	0.96	0.96	0.97
Cis Only	0.98	0.97	0.97	0.97

Table 45 This table reports the p-values for chi-square tests used to test for significant differences between the number of autosomal genes in each transcription category between males and females of each cross.

Category	ARAB x COLZ	COLZ x ARAB	QUAD x COLZ	COLZ x QUAD
Conserved	0.01	0.00	0.00	0.00
Compensatory	0.00	0.00	0.00	0.00
Trans Only	0.00	0.00	0.00	0.00
Cis x Trans	0.00	0.00	0.00	0.00
Cis + Trans	0.57	0.49	0.01	0.00
Cis Only	0.18	0.47	0.01	0.00

Table 46 Raw sequencing reads resulting from the sequencing effort of CQxQ male and female libraries and *An. coluzzi* (COLZ) and *An. quadriannulatus* (QUAD) parental strains.

Library Prep	Sequencing Strategy	Raw DNA Sequencing Reads
CQxQ Female 1	2 Lanes Illumina HiSeq 2500 Rapid (125 bp SE) + 1 Lane Illumina HiSeq 2500 High Throughput (125 bp SE)	24,329,816
CQxQ Female 2		20,505,754
CQxQ Female 3		24,152,111
CQxQ Female 4		27,100,826
CQxQ Female 5		33,405,791
CQxQ Male 1A		19,763,223
CQxQ Male 1B		23,776,981
CQxQ Male 1C		24,346,712
CQxQ Male 2		27,460,228
CQxQ Male 3		26,065,505
CQxQ Male 4		29,134,128
COLZ 1	1 Lane Illumina HiSeq 2500 High Throughput (125 bp PE)	41,718,756
COLZ 2		48,522,452
QUAD 1		45,776,218
QUAD 1		46,667,555

Table 47 Results of the multiple interval mapping analysis for the simple interval mapping three locus model. QTL locations and interactions are listed under the Effect heading.

	df	SS	MS	LOD	%var	Pvalue(Chi2)	Pvalue(F)	
Model	7	1395.69	199.38	89.64	62.49	0	0	
Error	413	837.82	2.03					
Total	420	2233.51						

Effect	df	Type III SS	LOD	%var	F - value	Pvalue(Chi2)	Pvalue(F)	
X@0.0	4	1202.60	81.37	53.84	148.20	0.00	< 2.00E-16	***
2@108.0	4	32.37	3.47	1.45	3.99	0.00	0.003461	**
3@101.4	4	8.78	0.95	0.39	1.08	0.36	0.365132	
X@0.0:2@108.0	2	28.22	3.03	1.26	6.96	0.00	0.001069	**
X@0.0:3@101.4	2	3.12	0.34	0.14	0.77	0.46	0.464404	
2@108.0:3@101.4	2	31.16	3.34	1.40	7.68	0.00	0.000531	***
X@0.0:2@108.0:3@101.4	1	22.37	2.41	1.00	11.03	0.00	0.000979	***

Table 48 Results of the multiple interval mapping analysis for the composite interval mapping three locus model. QTL locations and interactions are listed under the Effect heading.

	df	SS	MS	LOD	%var	Pvalue(Chi2)	Pvalue(F)	
Model	7	1386.66	198.09	88.66	62.08	0	0	
Error	413	846.85	2.05					
Total	420	2233.51						

Effect	df	Type III SS	LOD	%var	F - value	Pvalue(Chi2)	Pvalue(F)	
X@0.0	4	1216.49	81.41	54.47	148.32	0.00	< 2.00E-16	***
2@156.1	4	19.36	2.07	0.87	2.36	0.05	0.0527	.
3@98.0	4	-12.81	-1.39	-0.57	-1.56	1.00	1	
X@0.0:2@156.1	2	12.95	1.39	0.58	3.16	0.04	0.0436	*
X@0.0:3@98.0	2	-8.97	-0.97	-0.40	-2.19	1.00	1	
2@156.1:3@98.0	2	13.46	1.44	0.60	3.28	0.04	0.0385	*
X@0.0:2@156.1:3@98.0	1	12.12	1.30	0.54	5.91	0.01	0.0155	*

Table 49 Results of the multiple interval mapping analysis for the full, six locus Bayesian interval mapping model. The results from all QTL interactions were removed because they were not significant. QTL locations are listed under the Effect heading.

	df	SS	MS	LOD	%var	Pvalue(Chi2)	Pvalue(F)	
Model	63	1553.34	24.66	108.70	69.55	0	0	
Error	357	680.18	1.91					
Total	420	2233.51						

Effect	df	Type III SS	LOD	%var	F - value	Pvalue(Chi2)	Pvalue(F)	
X@0.0	32	1160.00	90.98	51.93	19.02	0.00	< 2.00E-16	***
2@72.7	32	71.43	9.13	3.20	1.17	0.11	0.24463	
2@102.1	32	56.05	7.24	2.51	0.92	0.40	0.5968	
2@131.7	32	62.95	8.09	2.82	1.03	0.24	0.42244	
2@159.9	32	64.93	8.34	2.91	1.07	0.20	0.37611	
3@101.4	32	105.50	13.18	4.72	1.73	0.00	0.00973	**

Table 50 Results of the multiple interval mapping analysis for the sub-setted four locus Bayesian interval mapping model. QTL locations and interactions are listed under the Effect heading.

	df	SS	MS	LOD	%var	Pvalue(Chi2)	Pvalue(F)
Model	15	1459.58	97.31	96.89	65.35	0	0
Error	405	773.93	1.91				
Total	420	2233.51					

Effect	df	Type III SS	LOD	%var	F - value	Pvalue(Chi2)	Pvalue(F)	
X@0.0	8	1172.32	84.30	52.49	76.69	0.00	< 2.00E-16	***
2@72.7	8	72.48	8.18	3.25	4.74	0.00	1.41E-05	***
2@159.9	8	58.92	6.71	2.64	3.85	0.00	0.000217	***
3@101.4	8	47.12	5.40	2.11	3.08	0.00	0.002195	**
X@0.0:2@72.7	4	39.78	4.58	1.78	5.21	0.00	0.000426	***
X@0.0:2@159.9	4	36.04	4.16	1.61	4.72	0.00	0.000993	***
X@0.0:3@101.4	4	27.01	3.14	1.21	3.53	0.01	0.007526	**
2@72.7:2@159.9	4	21.96	2.56	0.98	2.87	0.02	0.022821	*
2@72.7:3@101.4	4	58.85	6.70	2.63	7.70	0.00	5.49E-06	***
2@159.9:3@101.4	4	35.93	4.15	1.61	4.70	0.00	0.001018	**
X@0.0:2@72.7:2@159.9	2	21.12	2.46	0.95	5.53	0.00	0.004282	**
X@0.0:2@72.7:3@101.4	2	29.72	3.45	1.33	7.78	0.00	0.000485	***
X@0.0:2@159.9:3@101.4	2	32.73	3.79	1.47	8.56	0.00	0.000228	***
2@72.7:2@159.9:3@101.4	2	19.88	2.32	0.89	5.20	0.01	0.005875	**
X@0.0:2@72.7:2@159.9:3 @101.4	1	18.11	2.12	0.81	9.48	0.00	0.00222	**

Table 51 *An. quadriannulatus*, sex-biased genes that are mis-expressed betwen COLZ x QUAD male hybrids and parental strains, and fall under significant sterility QTL.

Gene ID	Chromosome	bp location	QTL	Vectorbase or PantherDB Gene Description	UniProt Gene Ontology	Drosophila Ortholog
AGAP004221	2R	52,059,089	2R	AT04489P-RELATED (PTHR21391:SF1)		ryder cup (R-CUP), presidents-cup (P-CUP), walker cup (WA-CUP), CG7634, **Female meiosis, germ-line development during embryogenesis**
AGAP004312	2R	54,380,035	2R			
AGAP004476	2R	56,758,291	2R	GH10249P (PTHR23506:SF20), amino acid transporter(PC00046)	integral component of plasma membrane [GO:0005887]; presynapse [GO:0098793]; serotonin transmembrane transporter activity [GO:0015222]; aminergic neurotransmitter loading into synaptic vesicle [GO:0015842]	Vesicular monoamine transporter (VMAT), packaging the neurotransmitters dopamine, serotonin and octopamine into secretory vesicles

AGAP005163	2L	10,907,702	2L (A)	glucosyl/glucuronosyl transferases [Source:VB Community Annotation]	integral component of membrane [GO:0016021]; transferase activity, transferring hexosyl groups [GO:0016758]; metabolic process [GO:0008152]	CG3797
AGAP005217	2L	12,418,732	2L (A)			CG6362
AGAP005340	2L	14,264,824	2L (A)			
AGAP005435	2L	15,466,771	2L (A)	IODOTYROSINE DEHALOGENASE 1 (PTHR23026:SF120), peroxidase(PC00180)	integral component of membrane [GO:0016021]; oxidoreductase activity [GO:0016491]	CG6279
AGAP005519	2L	16,587,327	2L (A)			

AGAP005756	2L	20,287,341	2L (A)	Carboxylic ester hydrolase, CARBOXYLIC ESTER HYDROLASE (PTHR11559:SF262), esterase(PC00097);lipase(PC00143)	carboxylic ester hydrolase activity [GO:0052689]; metabolic process [GO:0008152]	cricklet (CLT), **male mating behavior**
AGAP006170	2L	27,641,188	2L (B)	AT24369P (PTHR33588:SF2)		CG14017, CG14013, CG3528, CG8138
AGAP006400	2L	31,095,979	2L (B)	Alkaline phosphatase, AT01495P-RELATED (PTHR11596:SF67), nucleotide phosphatase(PC00173)	alkaline phosphatase activity [GO:0004035]	
AGAP006432	2L	31,789,002	2L (B)	**MAJOR SPERM PROTEIN (PTHR23301:SF39), CHITIN BINDING PERITROPHIN-A (PTHR23301)**	extracellular region [GO:0005576]; chitin binding [GO:0008061]; chitin metabolic process [GO:0006030]	

AGAP006637	2L	35,402,172	2L (B)	Solute carrier organic anion transporter family member, SOLUTE CARRIER ORGANIC ANION TRANSPORTER FAMILY MEMBER (PTHR11388:SF108), transporter(PC00227)	integral component of membrane [GO:0016021]; plasma membrane [GO:0005886]; transporter activity [GO:0005215]; ion transport [GO:0006811]	
AGAP006655	2L	35,757,396	2L (B)		integral component of membrane [GO:0016021]	CG31036
AGAP006795	2L	38,567,686	2L (B)	**Peritrophin-1, MAJOR SPERM PROTEIN (PTHR23301:SF39), CHITIN BINDING PERITROPHIN-A (PTHR23301)**	peritrophic matrix protein; midgut	
AGAP006934	2L	40,127,227	2L (B)			CG13924
AGAP007155	2L	43,592,241	2L (C)	825-OAK (PTHR35685:SF3)		
AGAP007156	2L	43,595,097	2L			

			(C)			
AGAP010887	3L	12,573,817	3L	CUTICULAR PROTEIN 47EE (PTHR10380:SF168)	structural constituent of cuticle [GO:0042302]	
AGAP010901	3L	12,800,494	3L	cuticular protein 2 from fifty-one aa family [Source:VB Community Annotation]		CG8543, CG7548, CG8541
AGAP010902	3L	12,810,781	3L	cuticular protein 1 from CPFL family [Source:VB Community Annotation]		
AGAP011032	3L	14,761,655	3L	RADIAL SPOKE HEAD PROTEIN 9 HOMOLOG (PTHR22069:SF3), ribosomal protein(PC00202)	**axoneme assembly [GO:0035082]; cilium movement involved in cell motility [GO:0060294]; motile cilium assembly [GO:0044458]**	CG31803

APPENDIX B

FIGURES

Chapter I

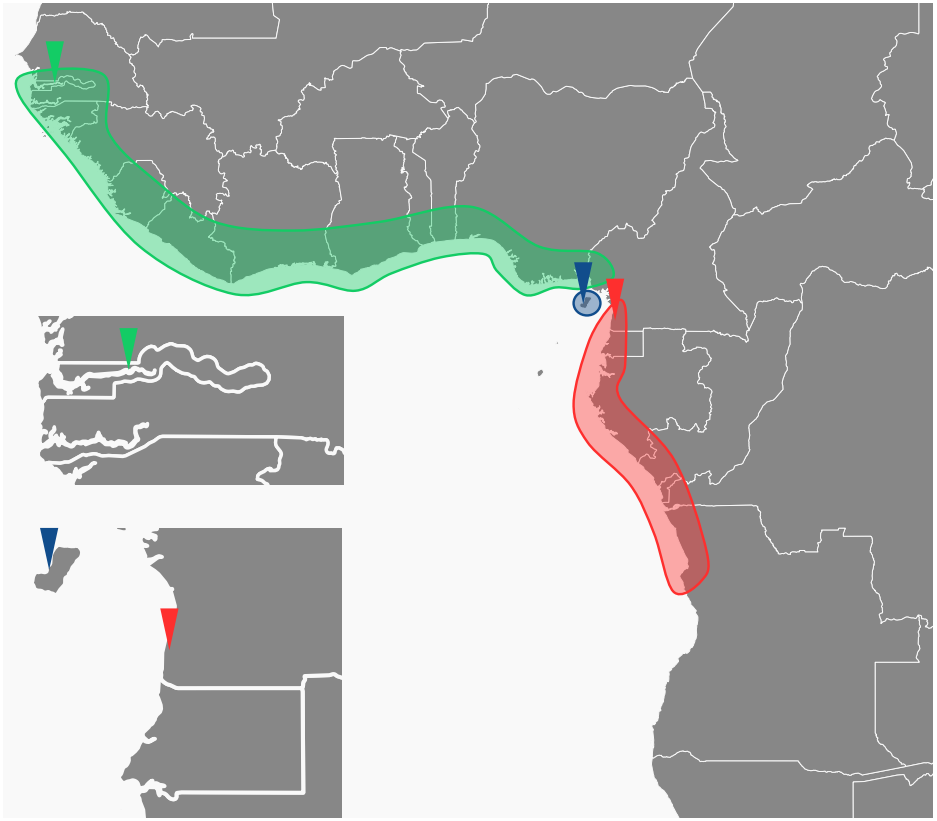


Figure 1 This map of West Africa illustrates the distributions of *An. melas* genetic clusters. Ranges of *An. melas* West (green), South (red), and Bioko (blue) are shown as shaded regions. Triangles show the sample locations of *An. melas* populations used to represent each *An. melas* genetic cluster. The top inset shows the collection location of Ballingho, The Gambia (green triangle, *An. melas* West), and the bottom inset shows the collection locations of Arena Blanca, Bioko Island, Equatorial Guinea (blue triangle, *An. melas* Bioko) and Ipono, Cameroon (red triangle, *An. melas* South. Reprinted from Deitz *et al.* (2016).

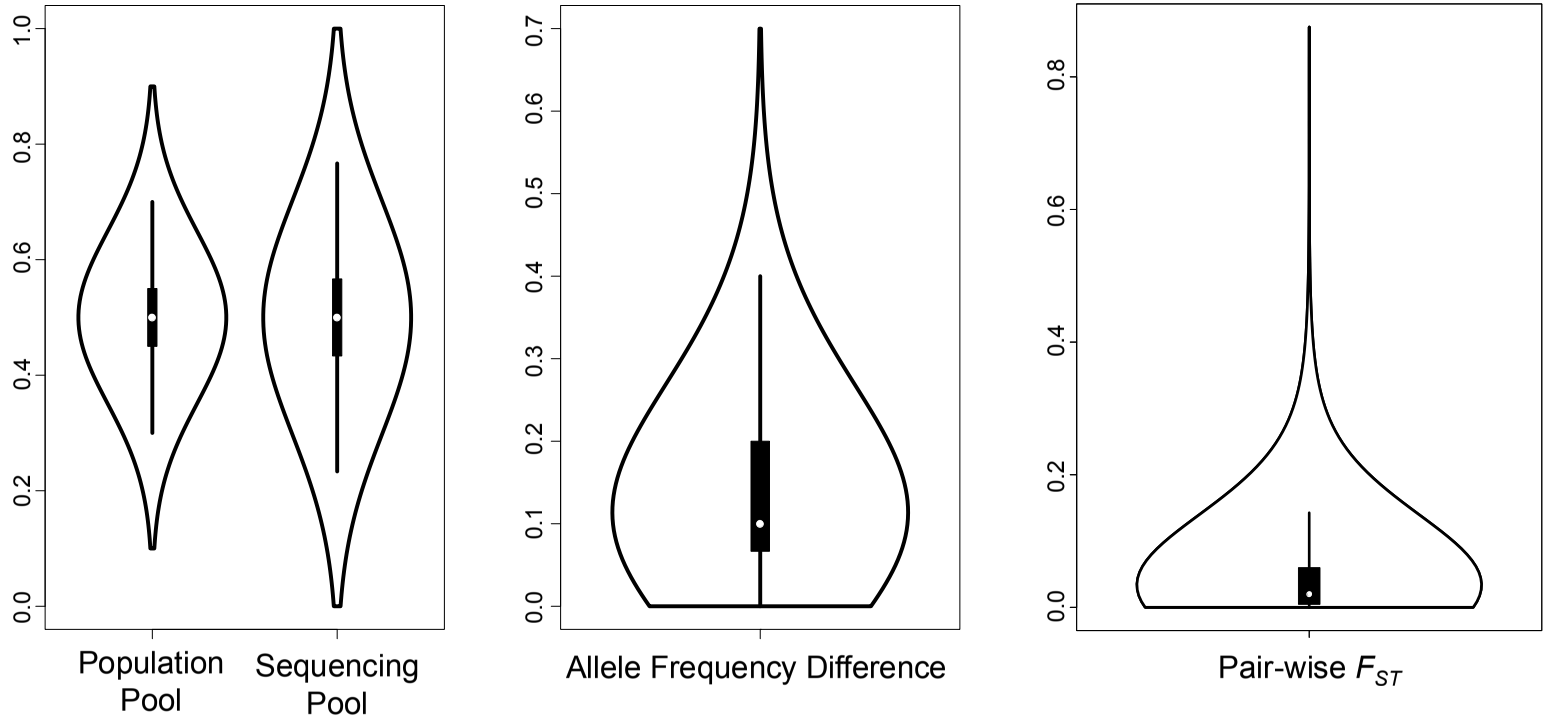
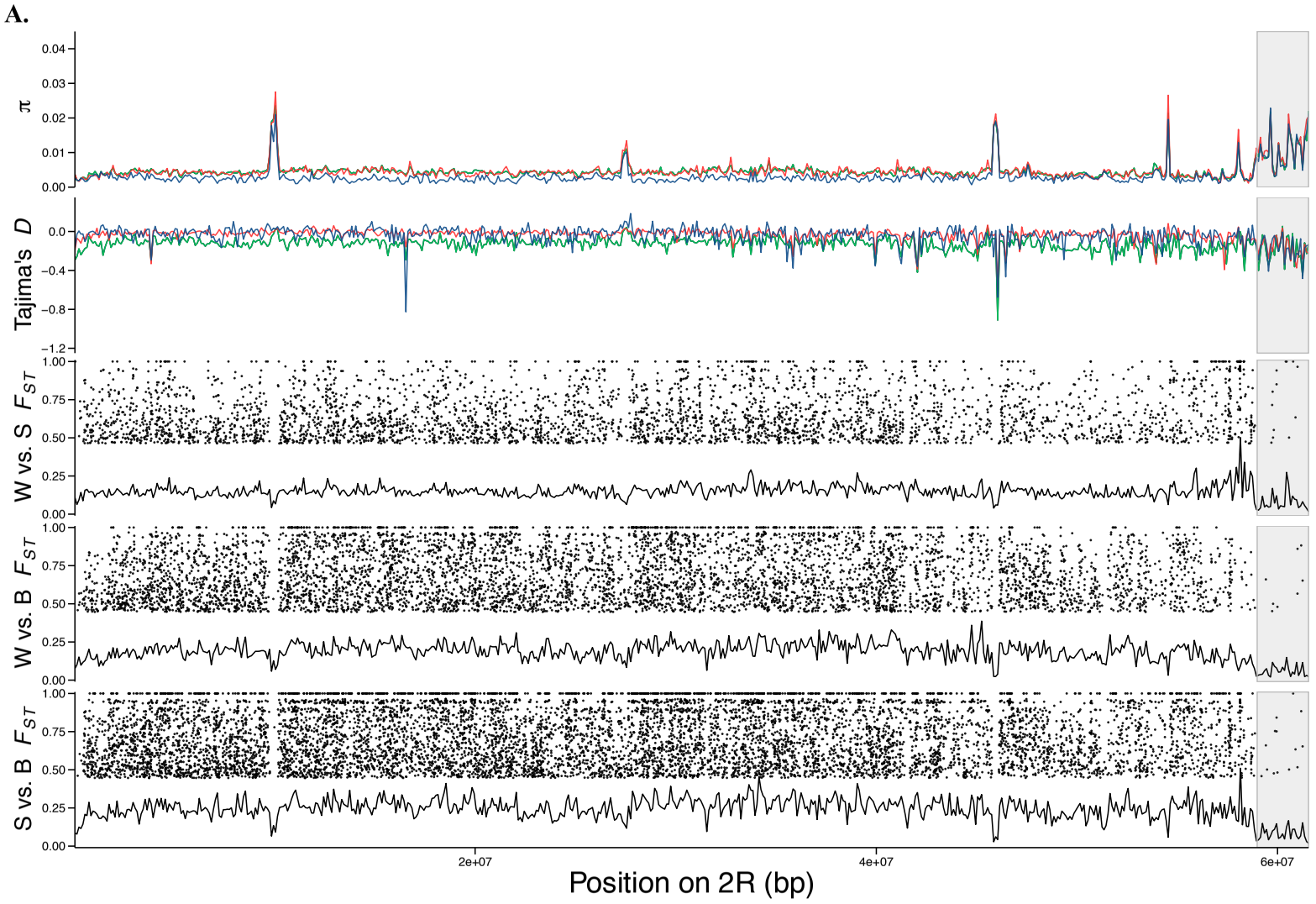
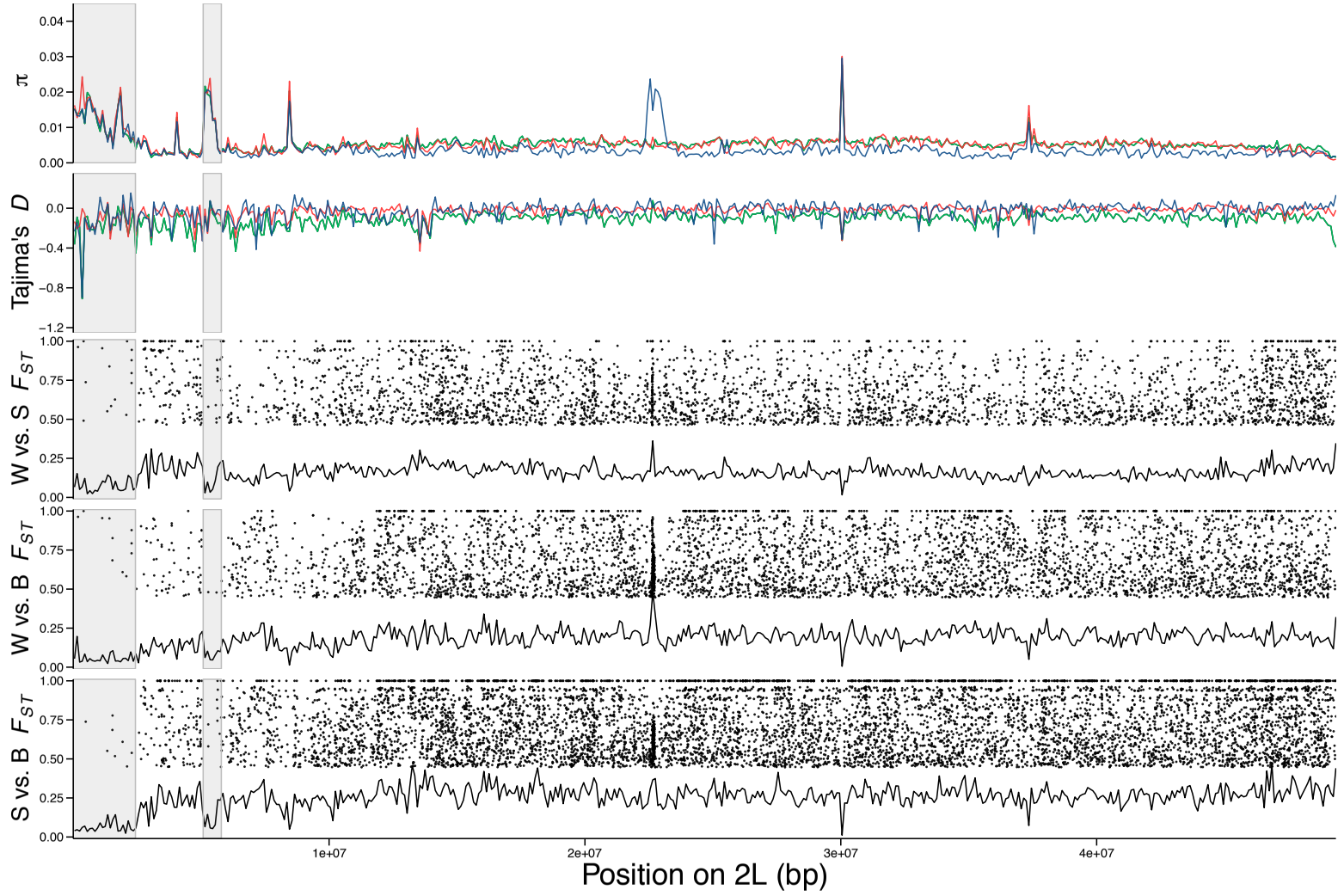


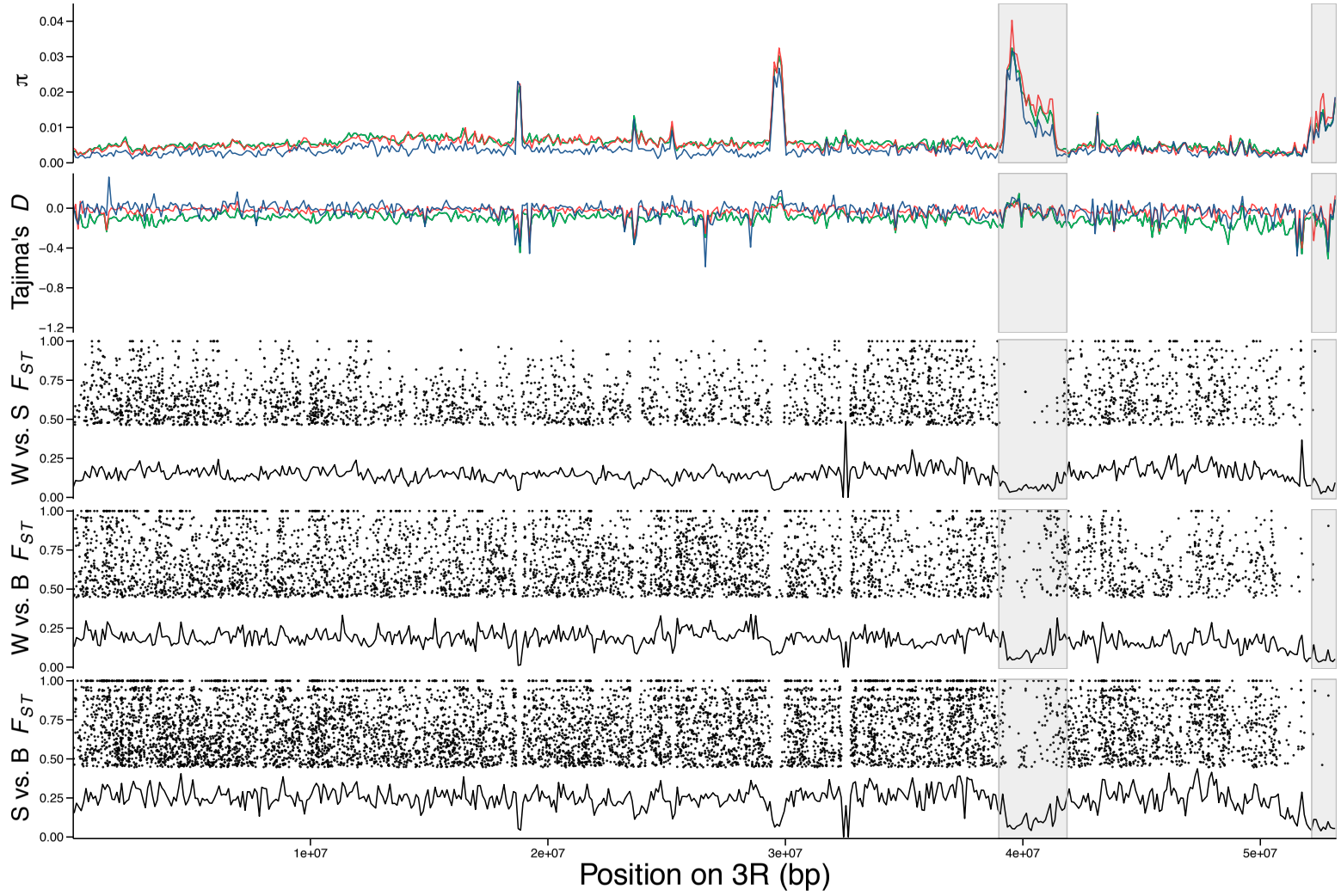
Figure 2 Summary violin plots of the F_{ST} null distribution and false discovery rate simulation. The left plots show the allele frequency distribution of population and sequencing pools. The middle plot represents the difference between two randomly sampled allele frequencies drawn from the sequencing pool. The right plot shows the distribution of F_{ST} values calculated from distribution of allele frequency difference. Reprinted from Deitz *et al.* (2016).


A.
π
0.04
0.03
0.02
0.01
0.00
Tajima's D
0.0
−0.4
−0.8
−1.2
W vs. S F_ST
W vs. B F_ST
S vs. B F_ST
1.00
0.75
0.50
0.25
0.00
2e+07
4e+07
6e+07
Position on 2R (bp)

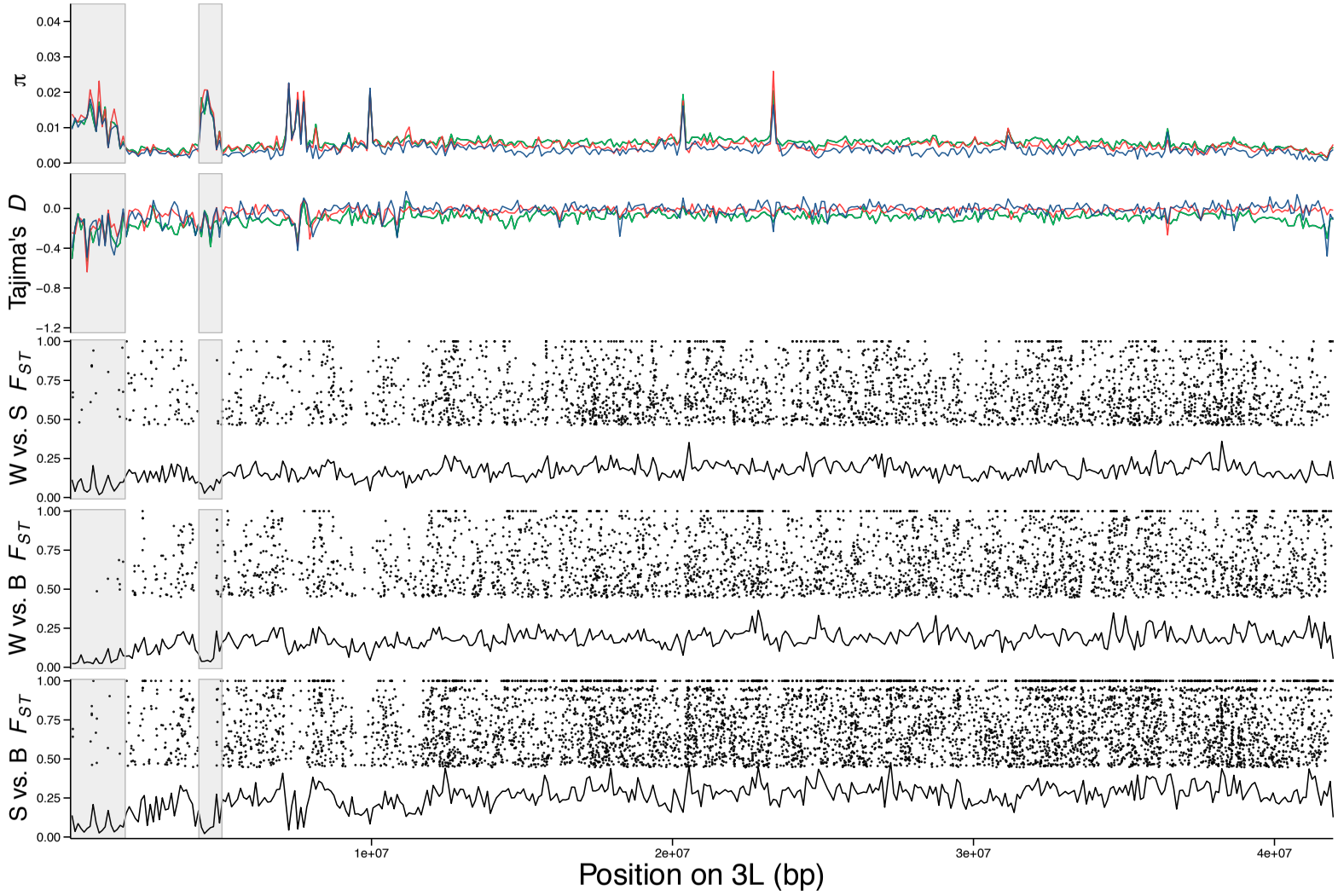
B.


π
0.04
0.03
0.02
0.01
0.00
Tajima's D
0.0
−0.4
−0.8
−1.2
W vs. S F_ST
W vs. B F_ST
S vs. B F_ST
1.00
0.75
0.50
0.25
0.00
1e+07
2e+07
3e+07
4e+07
Position on 2L (bp)

C.


π
Tajima's D
W vs. S F_ST
W vs. B F_ST
S vs. B F_ST
0.04
0.03
0.02
0.01
0.00
0.0
−0.4
−0.8
−1.2
1.00
0.75
0.50
0.25
0.00
1e+07
2e+07
3e+07
4e+07
5e+07
Position on 3R (bp)

D.


π
0.04
0.03
0.02
0.01
0.00
Tajima's D
0.0
−0.4
−0.8
−1.2
W vs. S F_{ST}
W vs. B F_{ST}
S vs. B F_{ST}
1.00
0.75
0.50
0.25
0.00
1e+07
2e+07
3e+07
4e+07
Position on 3L (bp)

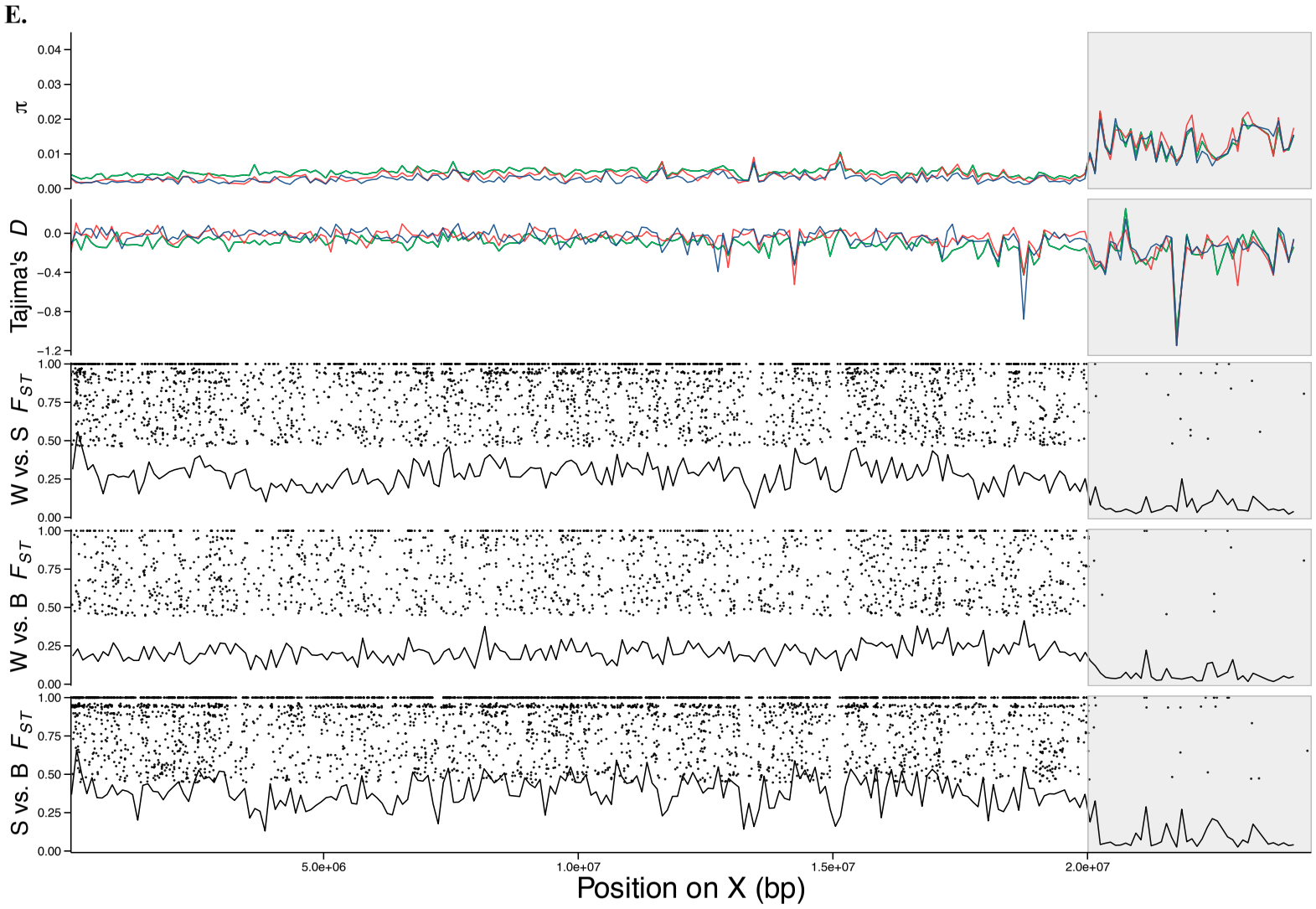

E.
π
0.04
0.03
0.02
0.01
0.00
Tajima's D
0.0
−0.4
−0.8
−1.2
W vs. S F_{ST}
W vs. B F_{ST}
S vs. B F_{ST}
1.00
0.75
0.50
0.25
0.00
5.0e+06
1.0e+07
1.5e+07
2.0e+07
Position on X (bp)

Figure 3 Line plots illustrate genome-wide nucleotide diversity (π) and Tajima's *D* estimates for each chromosome arm and population based upon a non-overlapping, 100 kb sliding window. Green lines represent *An. melas* West, red lines represent *An. melas* South, and blue lines represent *An. melas* Bioko. F_{ST} plots are presented for each pair-wise population comparison: *An. melas* West vs. South (W vs. S), West vs. Bioko (W vs. B), and South vs. Bioko (S vs. B). The solid line indicates F_{ST} calculated for non-overlapping, 100 kb sliding windows, and dots indicate significant F_{ST} SNPs. Vertical grey bars indicate regions of heterochromatin in the *An. gambiae* genome that were not included in the calculation of summary statistics. Panels present chromosome arms 2R (A), 2L (B), 3R (C), and 3L (D), and chromosome X (E). Reprinted from Deitz *et al.* (2016).

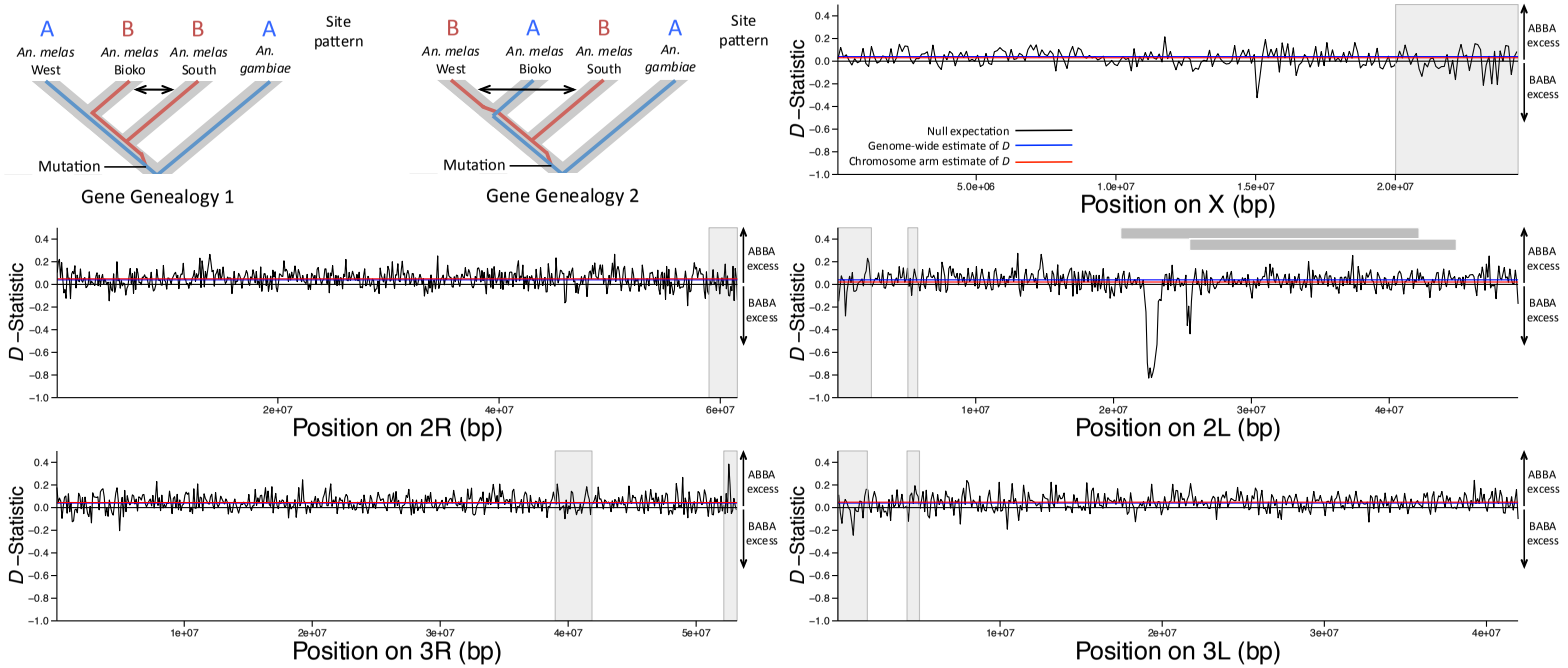


Figure 4 Line plots illustrate genome-wide values of Patterson's *D*-statistic for each chromosome arm for the *An. melas* population tree ((West,Bioko)South)*An. gambiae*). Positive values indicate an excess of ABBA patterns and negative values indicate a biased proportion of BABA patterns. Horizontal black lines indicate the null expectation, no ABBA or BABA excess ($D = 0$). Horizontal blue lines indicate the genome-wide estimate of Patterson's *D*, and horizontal red lines indicate the average Patterson's *D* for each chromosome arm. Vertical grey bars indicate regions of heterochromatin in the *An. gambiae* genome that were not included in the calculation of summary statistics. Horizontal grey bars in the chromosome arm 2L panel indicate the locations of the 2La/+ (top) and $2La^2/+$ (bottom) inversions. The top left panel demonstrates the ABBA vs. BABA patterns in the context of the *An. melas* tree, where an ABBA patter indicates introgression between *An. melas* Bioko and South, and a BABA pattern indicates introgression between *An. melas* West and South (arrows). Reprinted from Deitz *et al.* (2016).

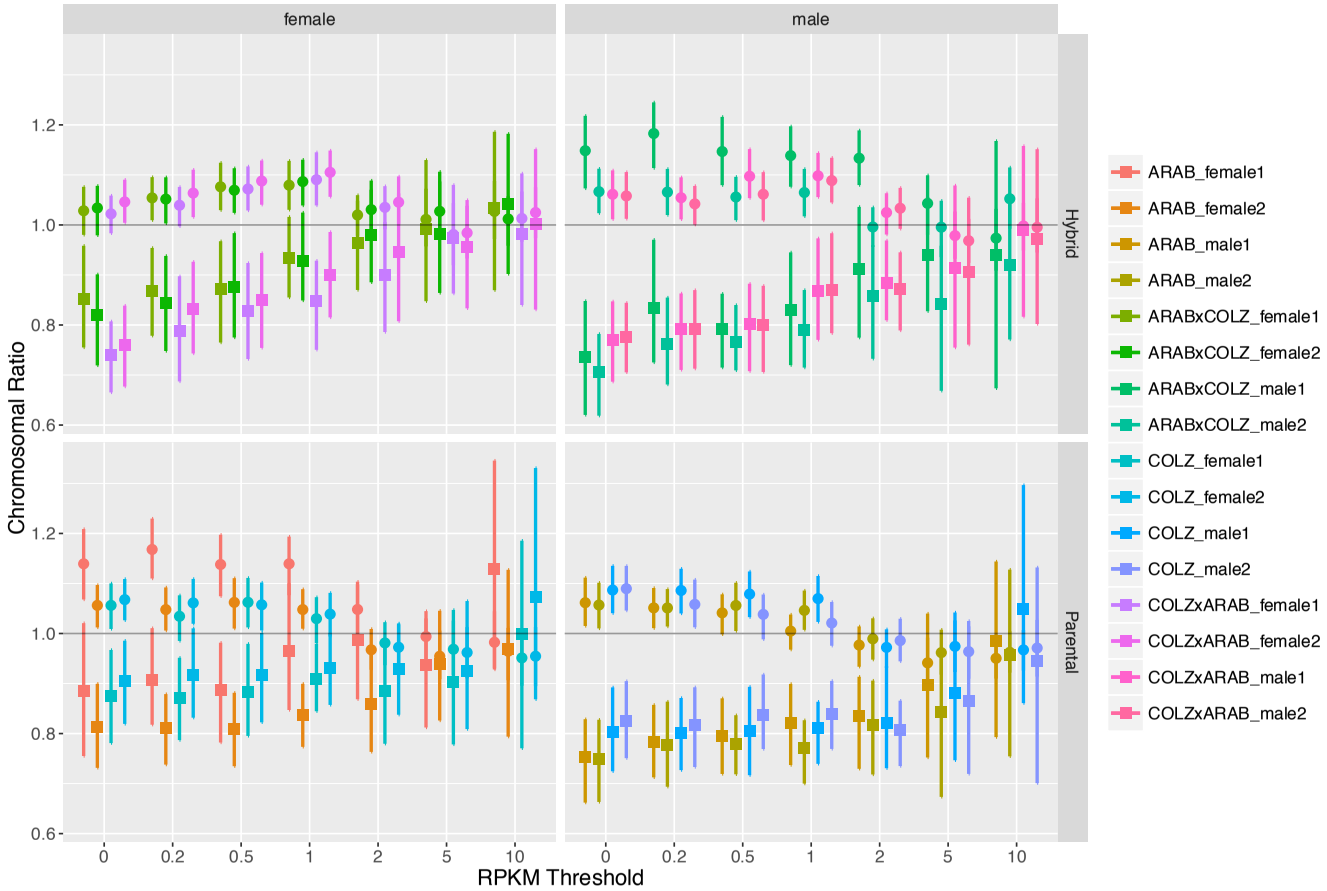


Figure 5 Scatter plot of X:A (squares) and 2:3 (circles) expression ratios at increasing minimum RPKM thresholds for the *An. coluzzii* - *An. arabiensis* species comparisons. Panels separate male and female, and parental and hybrid samples. Lines represent the 95% confidence interval of the median for each distribution.

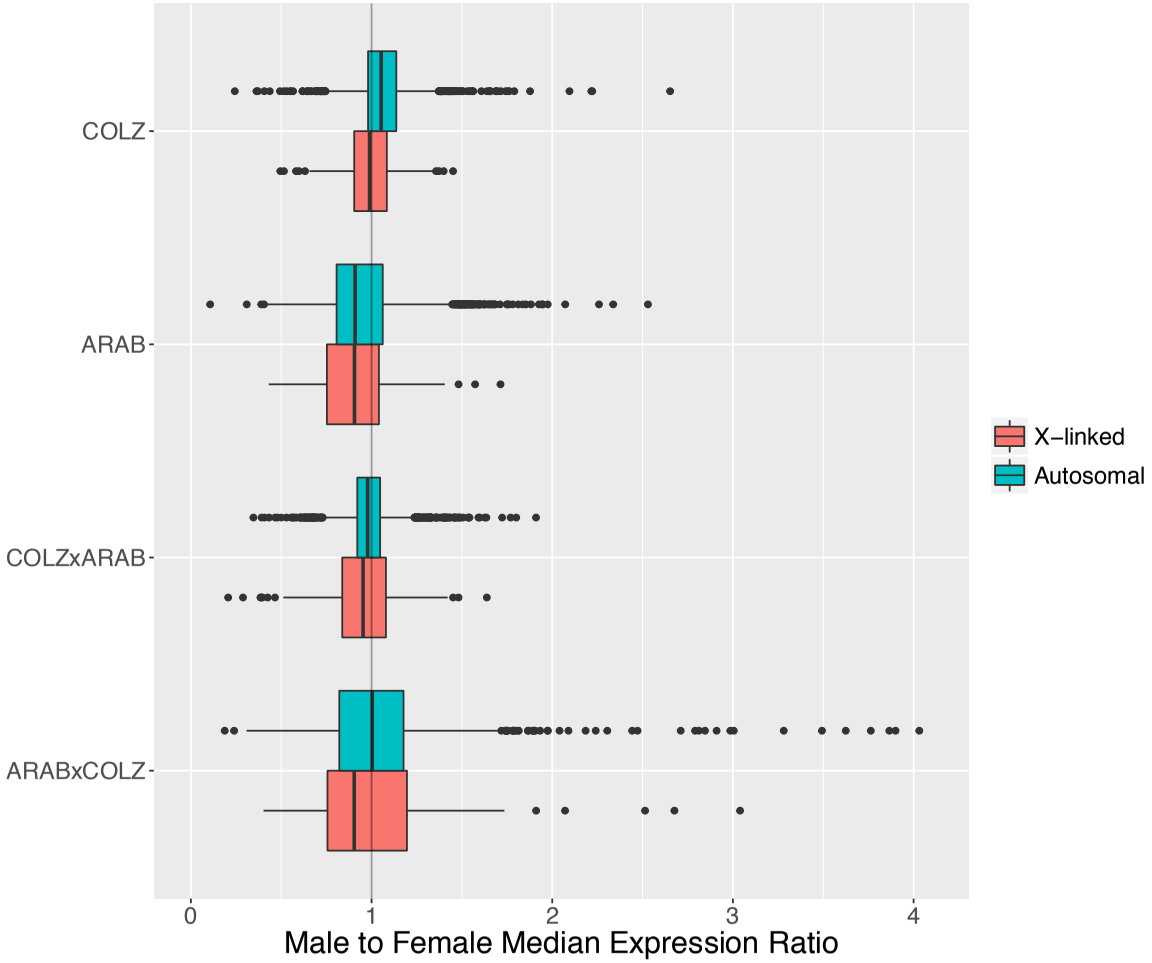


Figure 6 Box plot of M:F expression ratios distributions for the *An. coluzzii* - *An. arabiensis* species comparisons, separated by X-linked and autosomal genes.

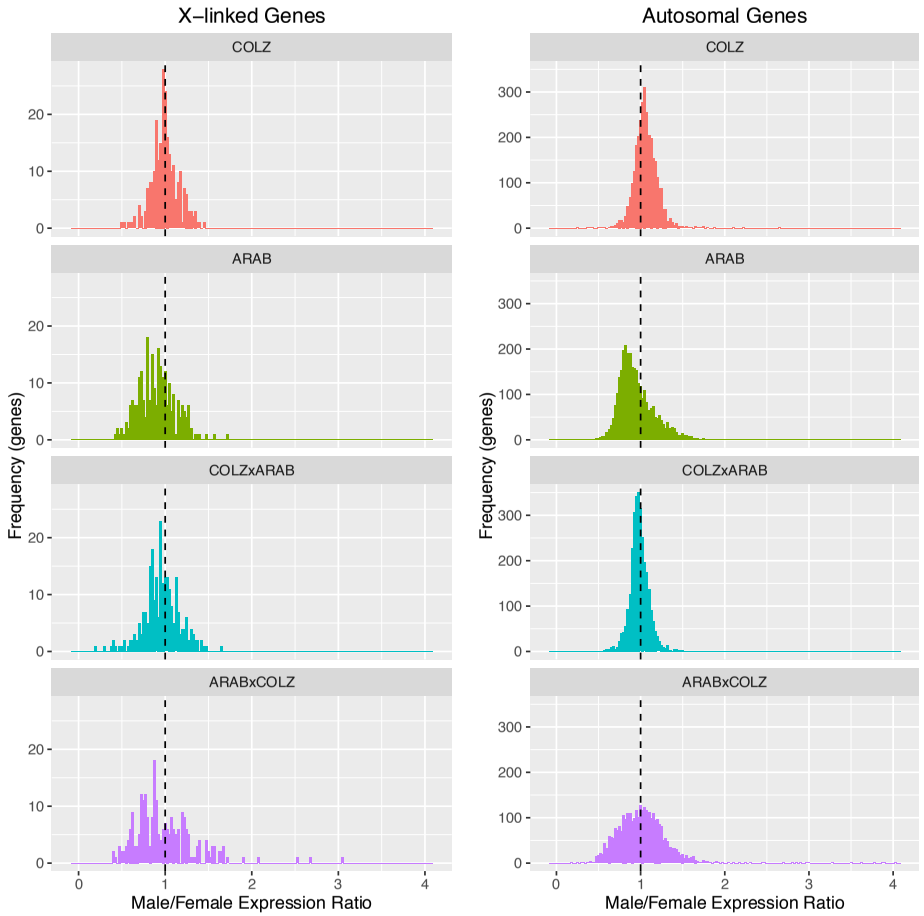


Figure 7 Histograms of the X-linked (left panel) and autosomal (right panel) M:F expression ratios distributions for the *An. coluzzii* - *An. arabiensis* species comparisons. Values above 1.0 (dotted line) indicate male-biased genes, while those below 1.0 indicate female-biased genes.

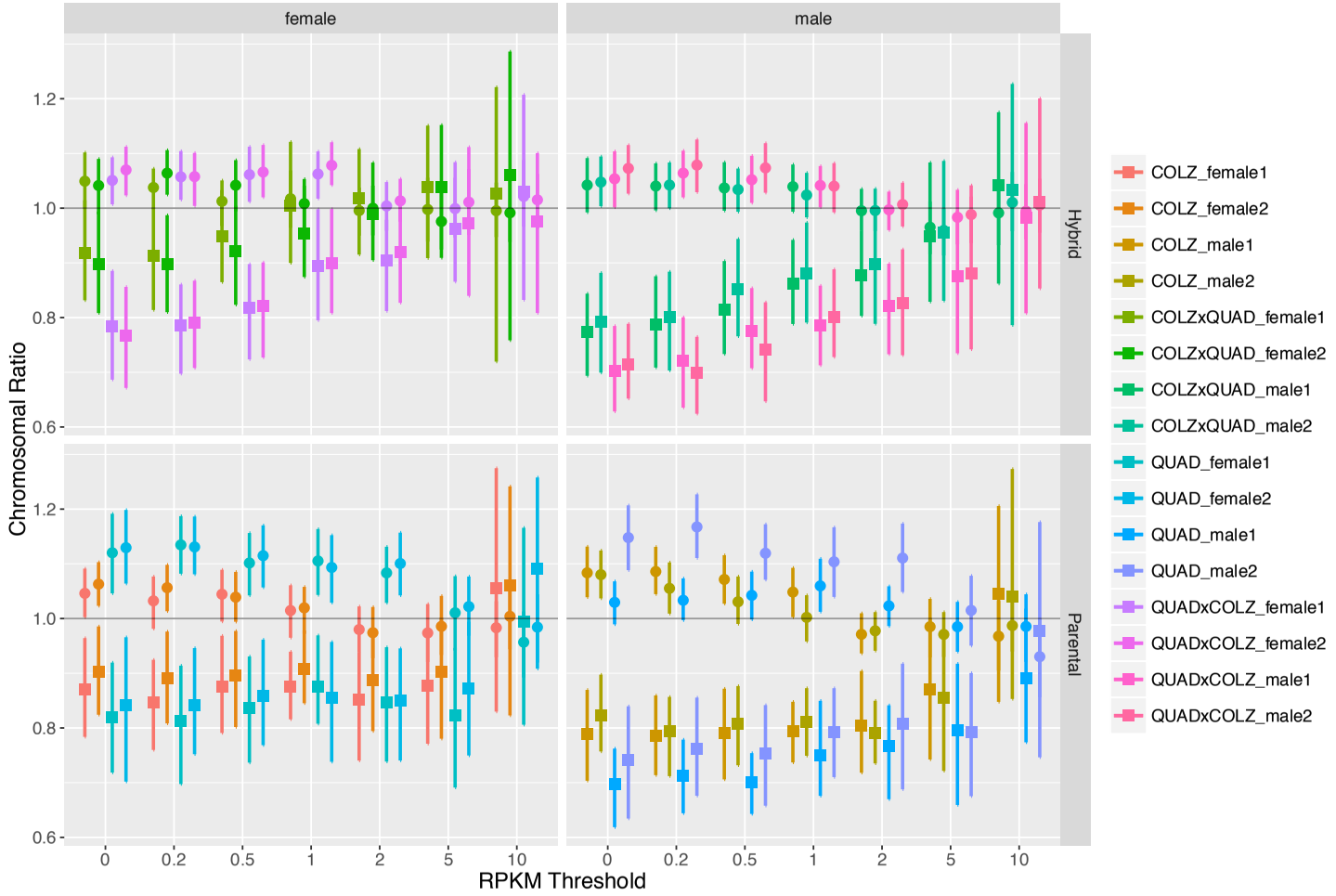


Figure 8 Scatter plot of X:A (squares) and 2:3 (circles) expression ratios at increasing minimum RPKM thresholds for the *An. coluzzii* - *An. quadriannulatus* species comparisons. Panels separate male and female, and parental and hybrid samples. Lines represent the 95% confidence interval of the median for each distribution.

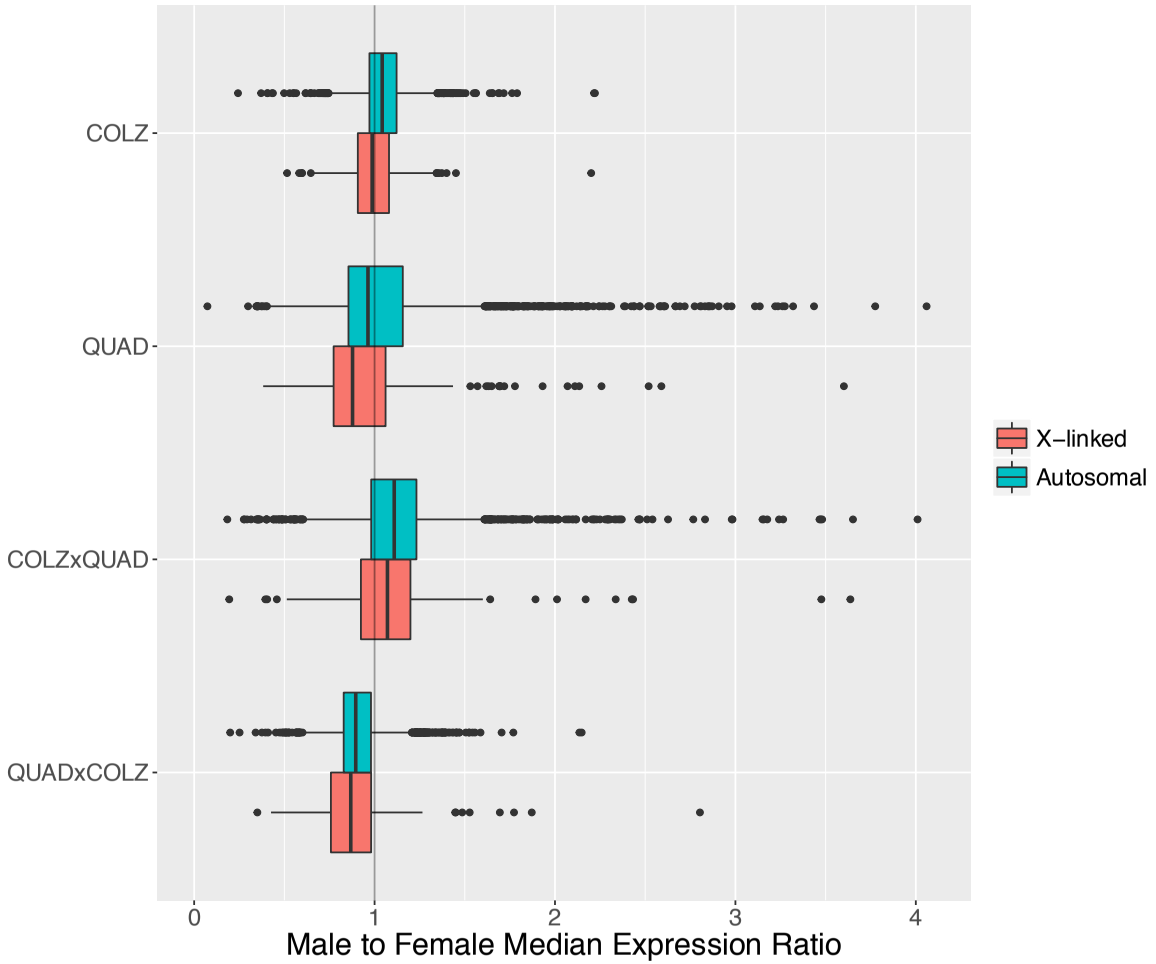


Figure 9 Box plot of M:F expression ratios distributions for the *An. coluzzii* - *An. quadriannulatus* species comparisons, separated by X-linked and autosomal genes.

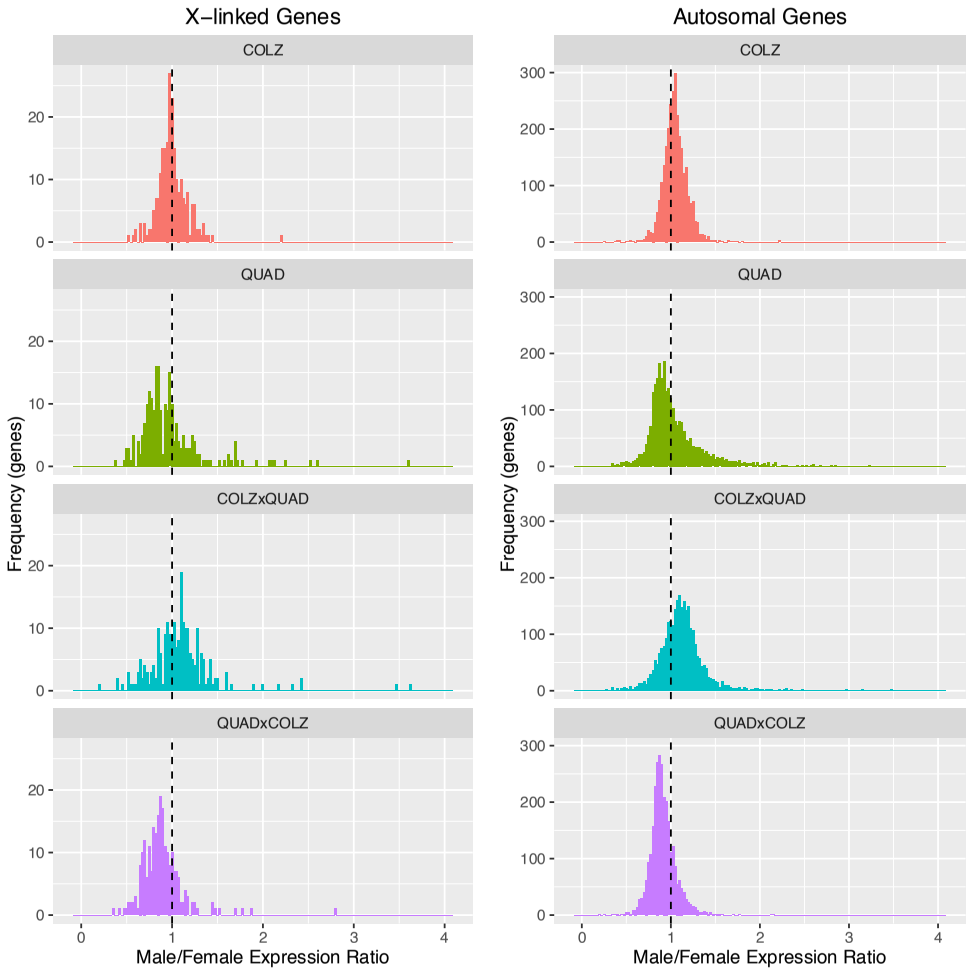


Figure 10 Histograms of the X-linked (left panel) and autosomal (right panel) M:F expression ratios distributions for the *An. coluzzii* - *An. quadriannulatus* species comparisons. Values above 1.0 (dotted line) indicate male-biased genes, while those below 1.0 indicate female-biased genes.

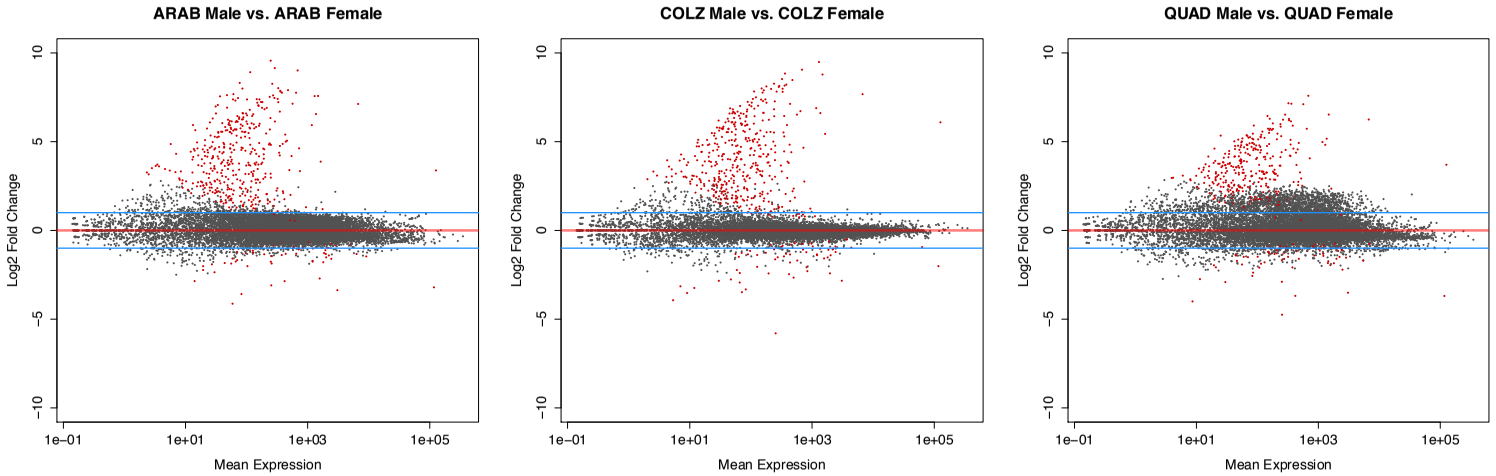


Figure 11 MA plots of gene expression differences between male and female *An. arabiensis* (ARAB, left), *An. coluzzii* (COLZ, center), and *An. quadriannulatus* (QUAD, right). Each dot represents a gene. The x-axis represents mean log intensity of the gene expression, and the y-axis represents the log2 fold change of the gene between males (positive) and females (negative). Red dots indicate genes that are have significantly different expression (p-adj. < 0.05) between males and females, while grey are not significantly different. Horizontal blue lines indicate the log2 fold change >1 and < -1 cutoffs to determine male- and female-biased gene expression, respectively.

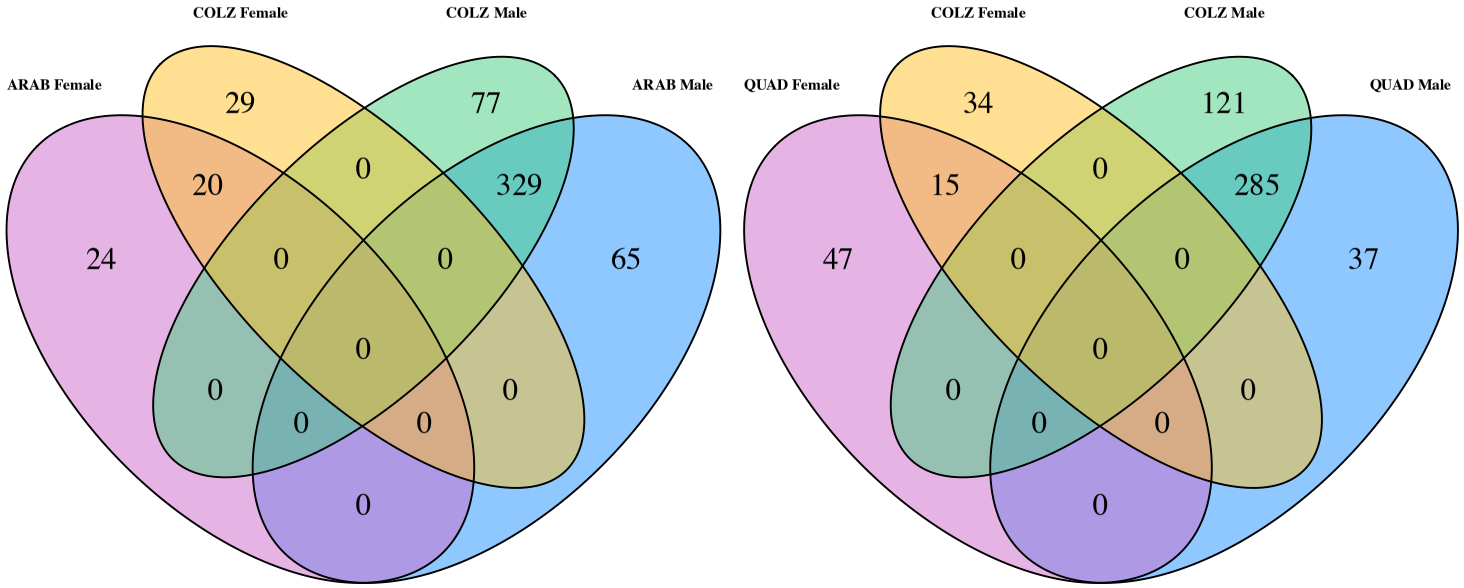


Figure 12 Venn diagram illustrate the overlap of male- and female-biased genes in each species pair used to generate F1 hybrids.

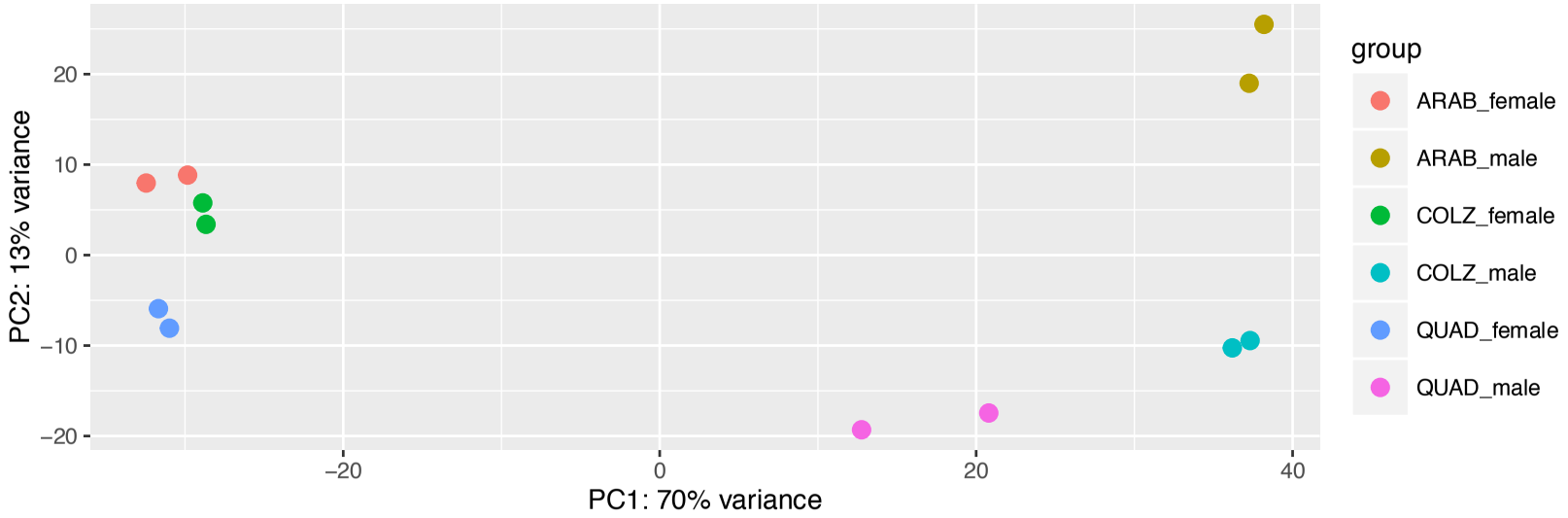


Figure 13 The results of a principal component analysis of the genome-wide expression and male and female *An. coluzzii* (COLZ), *An. arabiensis* (ARAB), and *An. quadriannulatus* (QUAD). The x and y axis explain 83% of the variance in the parental expression dataset. Two biological replicates of each species and sex are represented by colored dots.

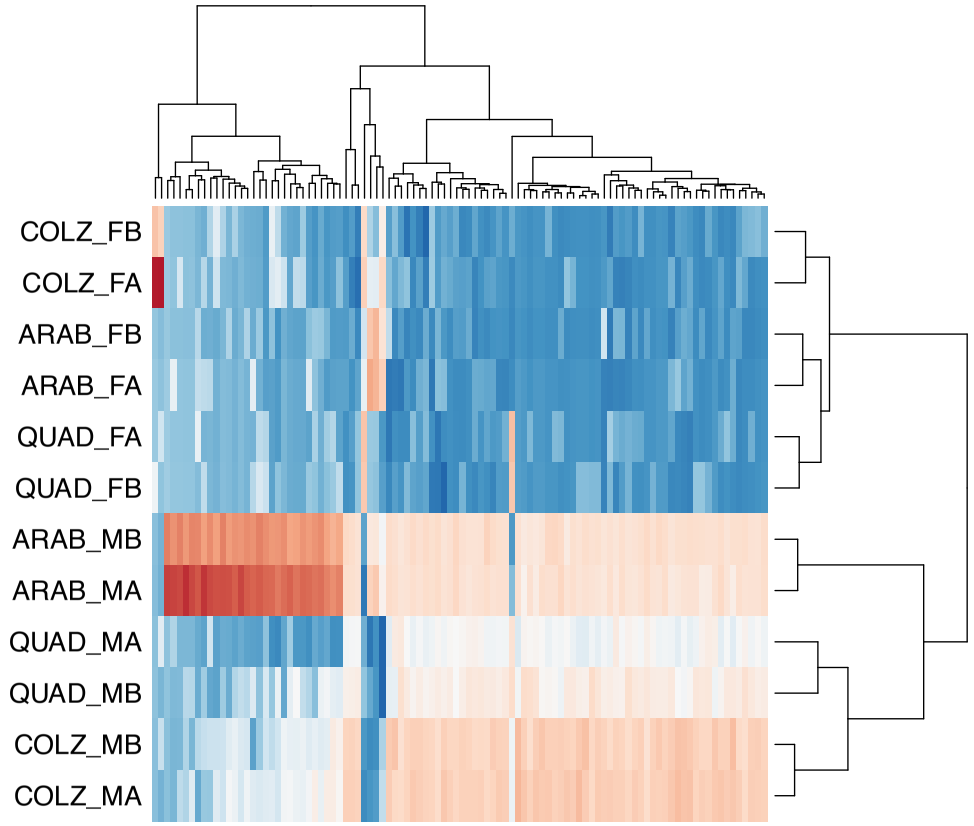


Figure 14 A heatmap of the 100 most variable, sex-specific genes from the combined set of all species (see Figure 12). Warm colors (red) higher relative gene expression, while cool colors (blue) indicate lower relative gene expression. The y-axis indicates how samples cluster according to expression similarity of these genes. Genes are clustered along the x-axis according to expression similarity. Biological replicates of COLZ, ARAB, and QUAD are annotated according to sex (M/F) and biological replicate (A/B). For example, COLZ_FA is *An. coluzzii* female A.

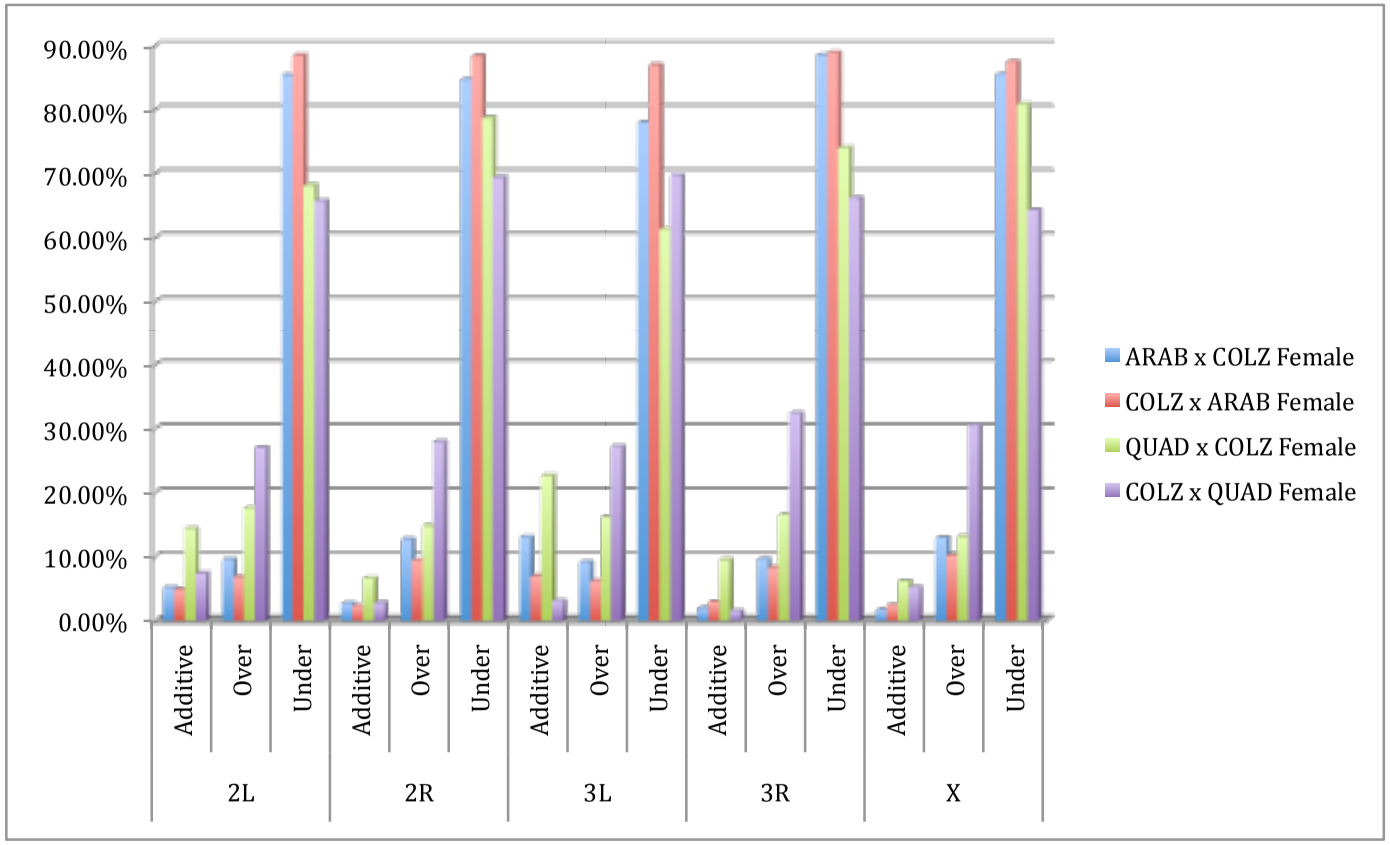


Figure 15 A histogram reporting the percentage of mis-expressed genes on each chromosome that fall into the categories of additive, over-, and under- expression for female F1 hybrids.

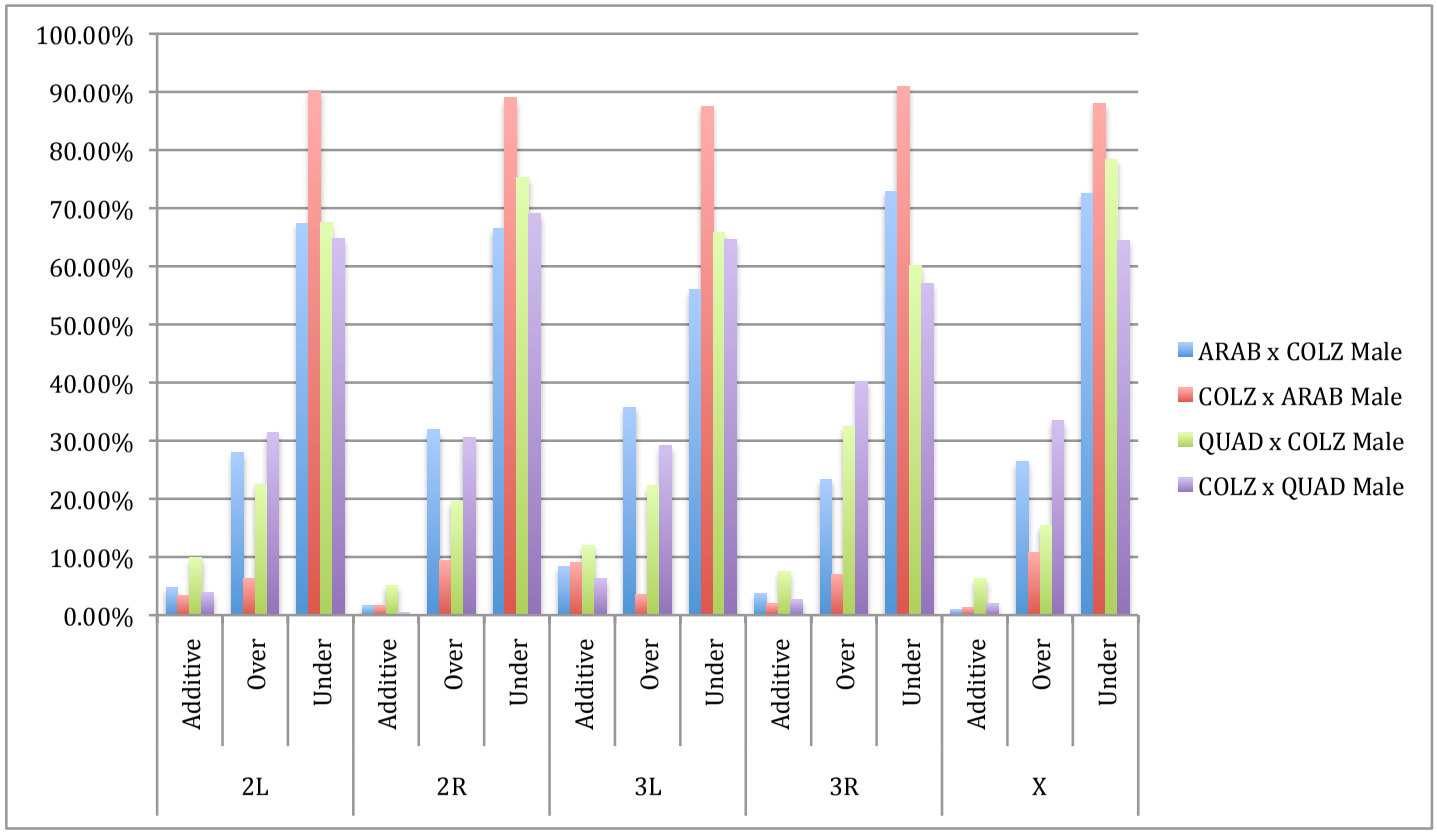


Figure 16 A histogram reporting the percentage of mis-expressed genes on each chromosome that fall into the categories of additive, over-, and under- expression for male F1 hybrids.

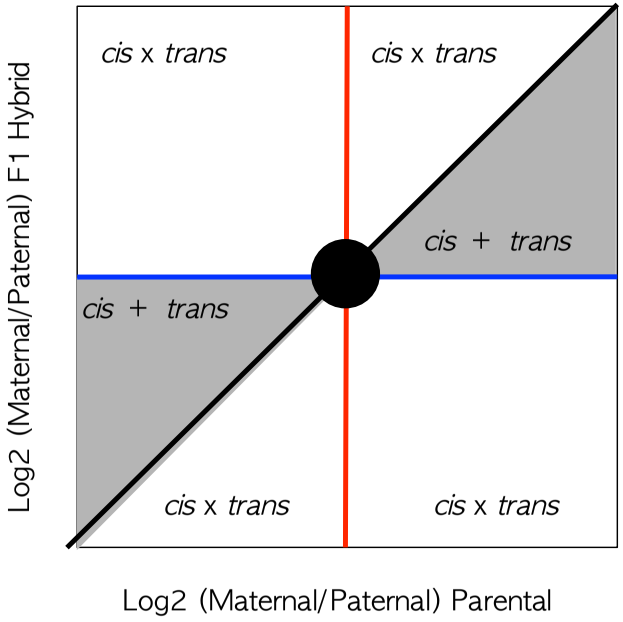


Figure 17
A scatter plot comparing maternal / paternal allele-specific expression ratios in hybrids (y-axis) to maternal / paternal expression ratios between parental strains (x-axis). Genes that are differentially expressed between parental alleles in hybrids will deviate from the horizontal line (y=0), and genes that are differentially expressed between parents will deviate from the vertical line (x=0). Genes that are not significantly differentiated between alleles in hybrid (not allelic imbalance) and are not significantly differentiated between parents are CONSERVED (black circle). Genes that are differentially expressed between alleles in a hybrid and/or between parental strains are categorized as: CIS ONLY (diagonal black line), TRANS ONLY (horizontal blue line), COMPENSATORY (vertical red line), CIS + TRANS (gray shaded), or CIS x TRANS (white shaded).

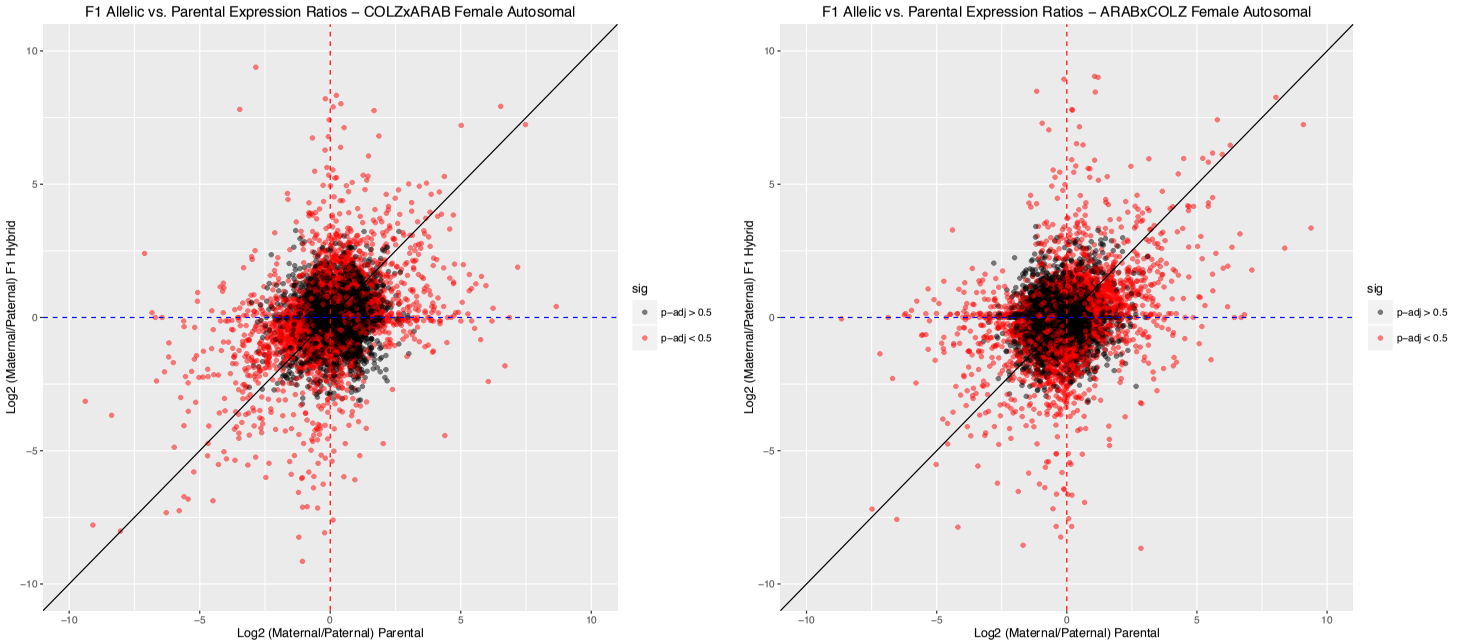


Figure 18 A scatter plot comparing maternal / paternal allele-specific expression ratios in hybrids (y-axis) to maternal / paternal expression ratios between parental strains (x-axis). The left graph represents the results of comparing allele-specific expression ratios in F1 **female** COLZ x ARAB **autosomal** genes to maternal (COLZ) / paternal (ARAB) expression ratios. The right graph represents the results of the opposite direction of the cross, comparing allele-specific expression ratios in F1 **female** ARAB x COLZ **autosomal** genes to maternal (ARAB) / paternal (COLZ) expression ratios. Genes that show significant allelic imbalance in hybrids and/or are differentially expressed between maternal / paternal species are plotted in red (p-adj < 0.05).

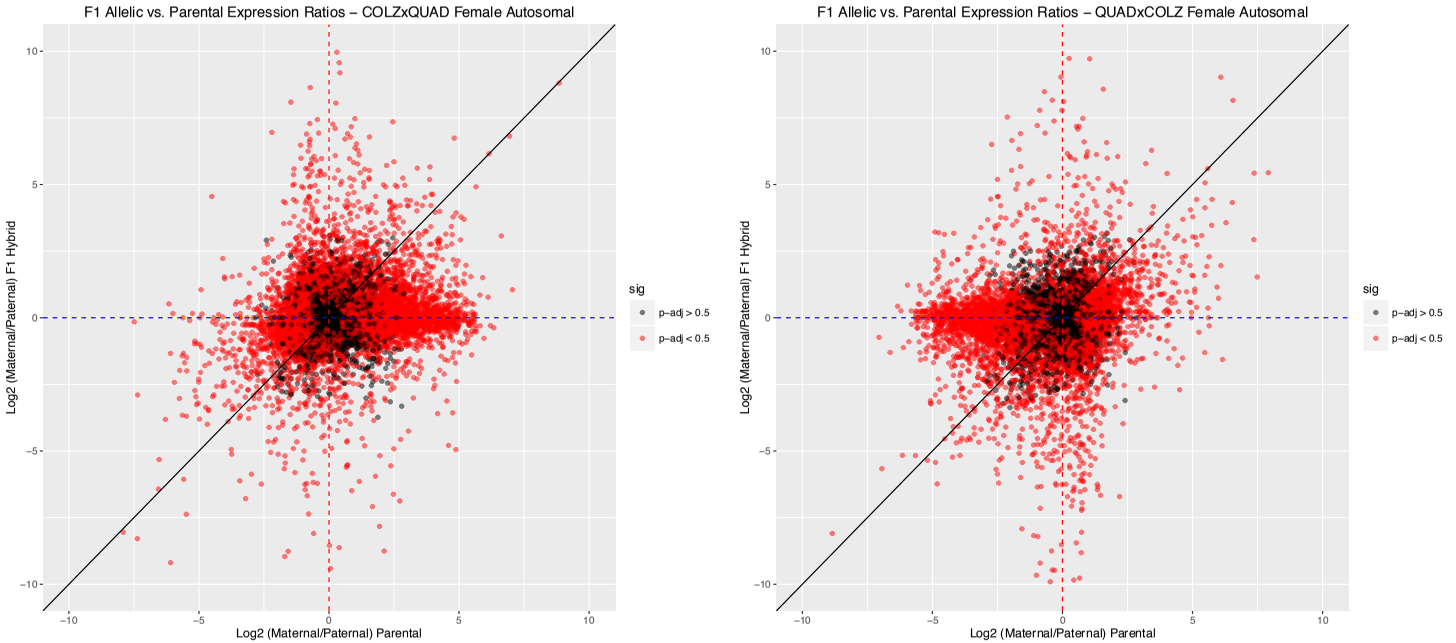


Figure 19 A scatter plot comparing maternal / paternal allele-specific expression ratios in hybrids (y-axis) to maternal / paternal expression ratios between parental strains (x-axis). The left graph represents the results of comparing allele-specific expression ratios in F1 **female** COLZ x QUAD **autosomal** genes to maternal (COLZ) / paternal (QUAD) expression ratios. The right graph represents the results of the opposite direction of the cross, comparing allele-specific expression ratios in F1 **female** QUAD x COLZ **autosomal** genes to maternal (QUAD) / paternal (COLZ) expression ratios. Genes that show significant allelic imbalance in hybrids and/or are differentially expressed between maternal / paternal species are plotted in red (p-adj < 0.05).

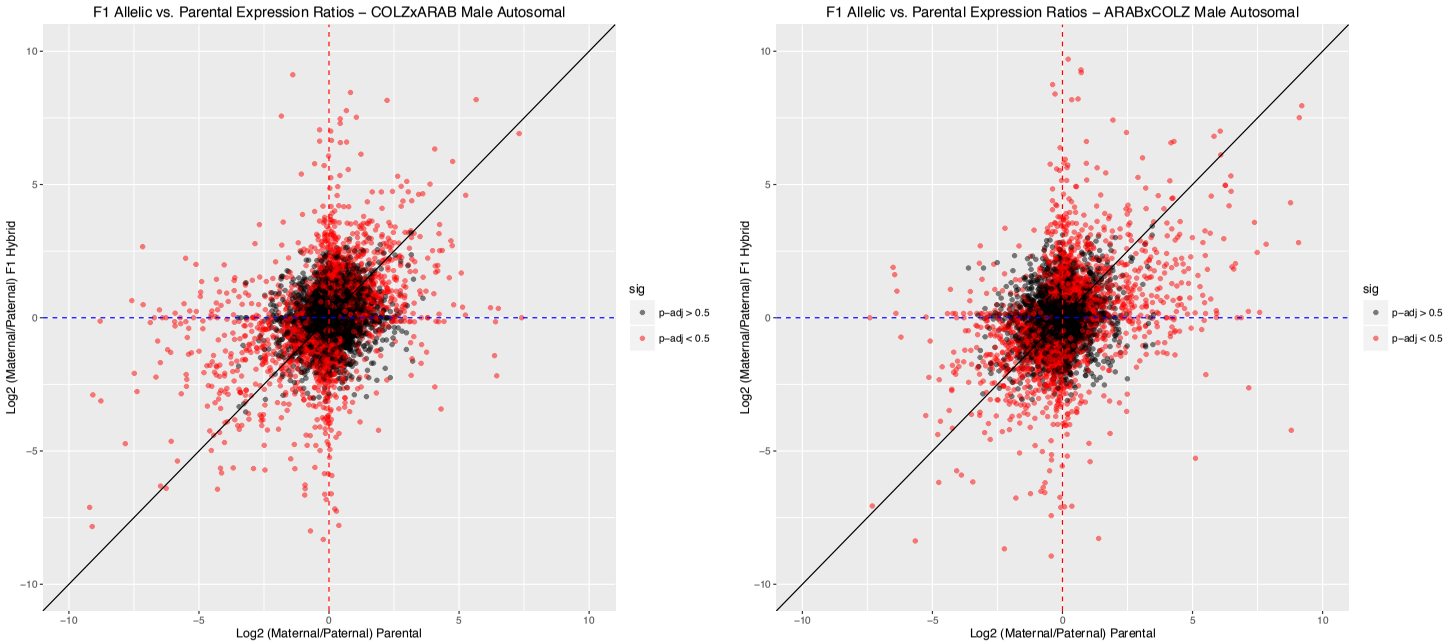


Figure 20 A scatter plot comparing maternal / paternal allele-specific expression ratios in hybrids (y-axis) to maternal / paternal expression ratios between parental strains (x-axis). The left graph represents the results of comparing allele-specific expression ratios in F1 **male** COLZ x ARAB **autosomal** genes to maternal (COLZ) / paternal (ARAB) expression ratios. The right graph represents the results of the opposite direction of the cross, comparing allele-specific expression ratios in F1 **male** ARAB x COLZ **autosomal** genes to maternal (ARAB) / paternal (COLZ) expression ratios. Genes that show significant allelic imbalance in hybrids and/or are differentially expressed between maternal / paternal species are plotted in red (p-adj < 0.05).

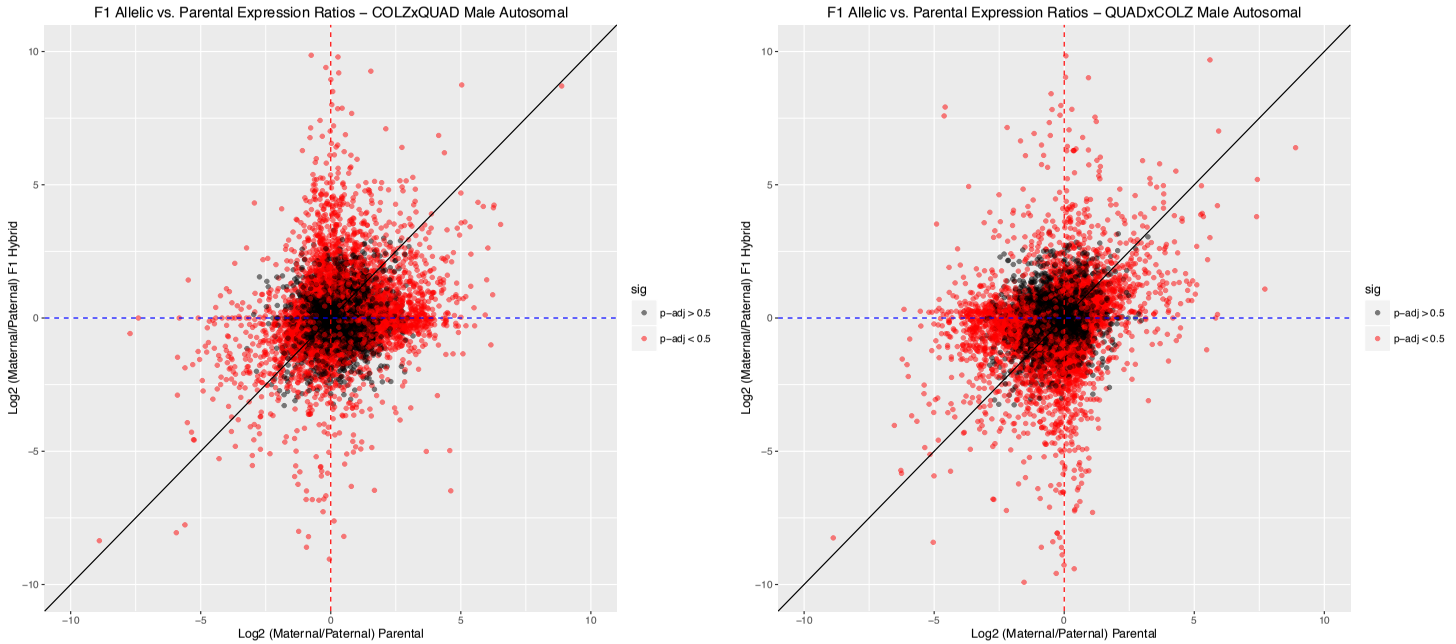


Figure 21 A scatter plot comparing maternal / paternal allele-specific expression ratios in hybrids (y-axis) to maternal / paternal expression ratios between parental strains (x-axis). The left graph represents the results of comparing allele-specific expression ratios in F1 **male** COLZ x QUAD **autosomal** genes to maternal (COLZ) / paternal (QUAD) expression ratios. The right graph represents the results of the opposite direction of the cross, comparing allele-specific expression ratios in F1 **male** QUAD x COLZ **autosomal** genes to maternal (QUAD) / paternal (COLZ) expression ratios. Genes that show significant allelic imbalance in hybrids and/or are differentially expressed between maternal / paternal species are plotted in red (p-adj < 0.05).

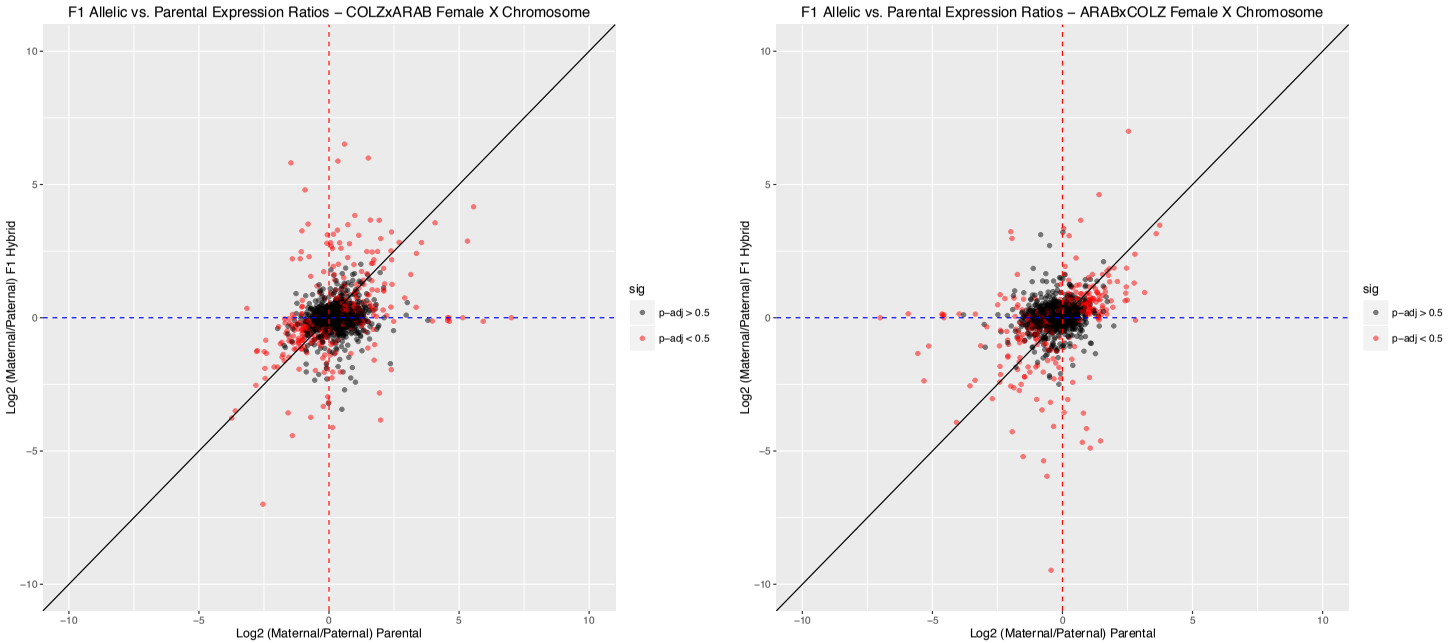


Figure 22 A scatter plot comparing maternal / paternal allele-specific expression ratios in hybrids (y-axis) to maternal / paternal expression ratios between parental strains (x-axis). The left graph represents the results of comparing allele-specific expression ratios in F1 **female** COLZ x ARAB **X chromosome** genes to maternal (COLZ) / paternal (ARAB) expression ratios. The right graph represents the results of the opposite direction of the cross, comparing allele-specific expression ratios in F1 **female** ARAB x COLZ **X chromosome** genes to maternal (ARAB) / paternal (COLZ) expression ratios. Genes that show significant allelic imbalance in hybrids and/or are differentially expressed between maternal / paternal species are plotted in red (p-adj < 0.05).

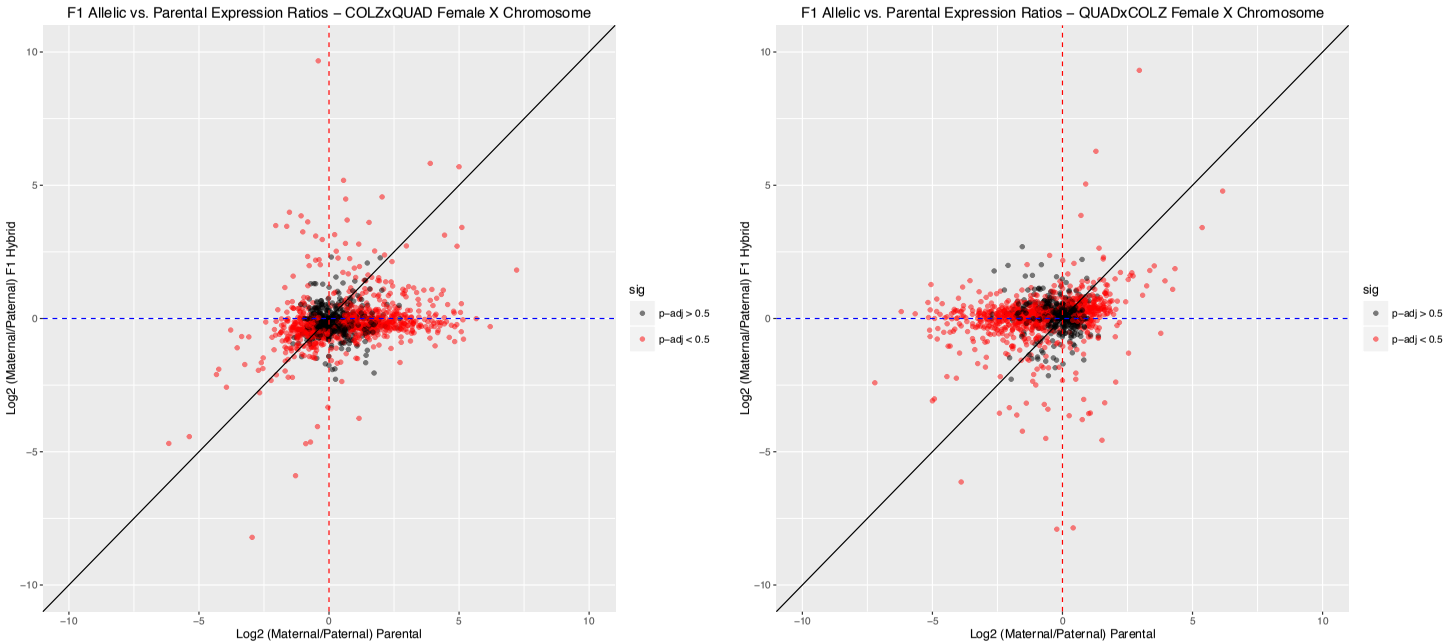


Figure 23 A scatter plot comparing maternal / paternal allele-specific expression ratios in hybrids (y-axis) to maternal / paternal expression ratios between parental strains (x-axis). The left graph represents the results of comparing allele-specific expression ratios in F1 **female** COLZ x QUAD **X chromosome** genes to maternal (COLZ) / paternal (QUAD) expression ratios. The right graph represents the results of the opposite direction of the cross, comparing allele-specific expression ratios in F1 **female** QUAD x COLZ **X chromosome** genes to maternal (QUAD) / paternal (COLZ) expression ratios. Genes that show significant allelic imbalance in hybrids and/or are differentially expressed between maternal / paternal species are plotted in red (p-adj < 0.05).

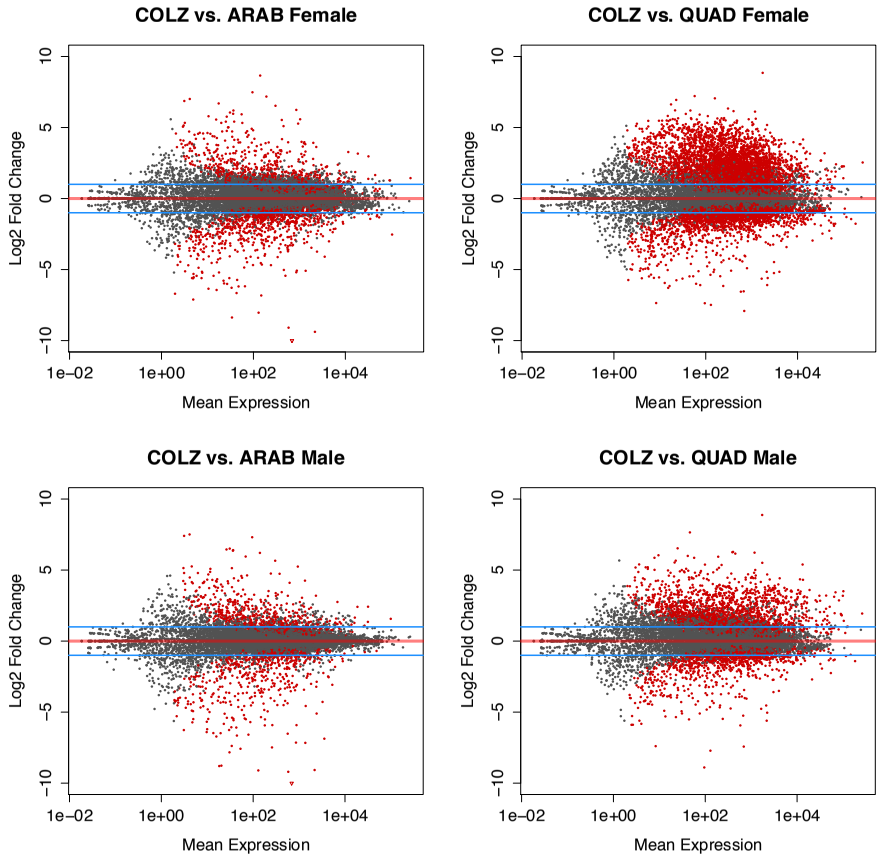


Figure 24 MA plots illustrate the relationship between intensity and difference between COLZ and ARAB females (top left) and males (bottom left), and COLZ and QUAD females (top right) and males (bottom right). The x-axis represents mean expression of the gene across biological replicates, and the y-axis represents log2 fold change of the gene. Positive values indicate COLZ-biased expression. A point represents each gene. Genes that are significantly differentiated between samples (p-adj. < 0.05) are shaded red.

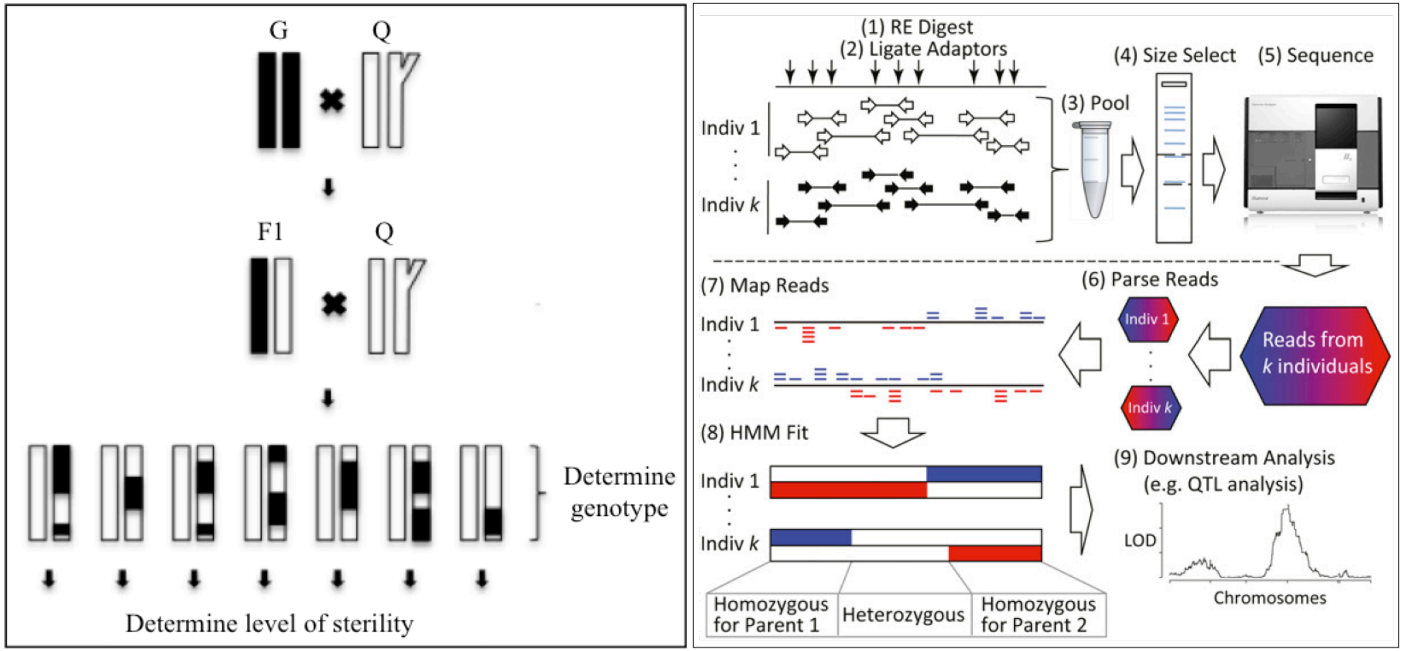


Figure 25 The left panel illustrates the backcross design used for QTL mapping in the CQxQ cross. The right panel illustrates the Multiplexed Shotgun Genotyping protocol used to genotype backcross individuals (Andolfatto *et al.,* 2010).

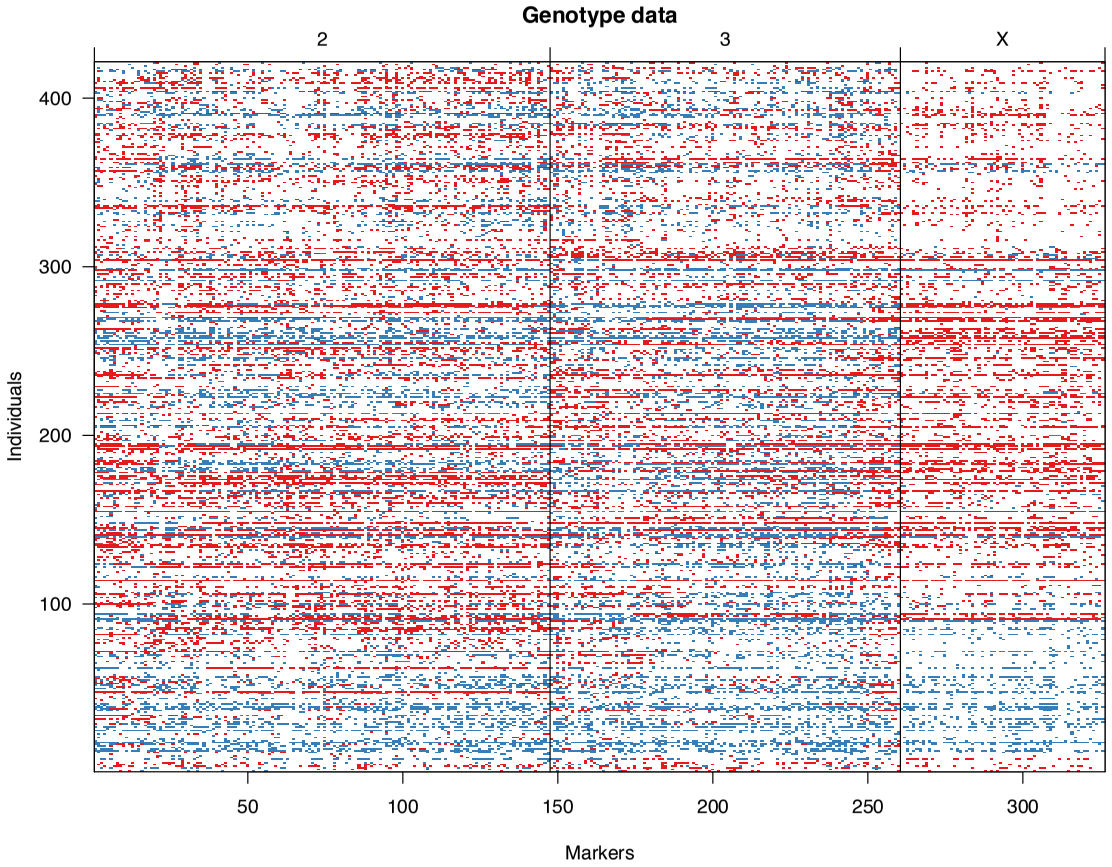


Figure 26. This figure illustrates the genotypes of each individual (y-axis) at all markers (x-axis) included in the QTL analysis. Homozygous *An. quadriannulatus* / *An. quadriannulatus* genotypes are blue, and heterozygous *An. coluzzi/An. quadriannulatus* genotypes are blue.

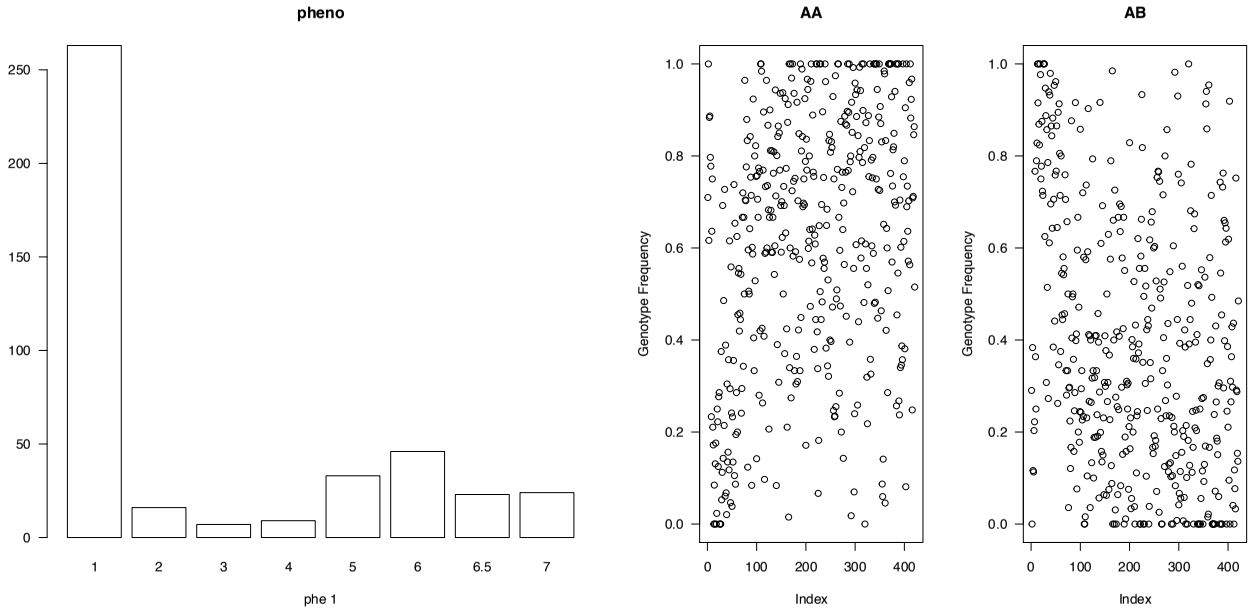


Figure 26 The left panel illustrates the distribution of the sterility phenotype observed in the in CQxQ males included in QTL mapping. The right panel illustrates the allele frequencies of the markers included in the QTL analysis.

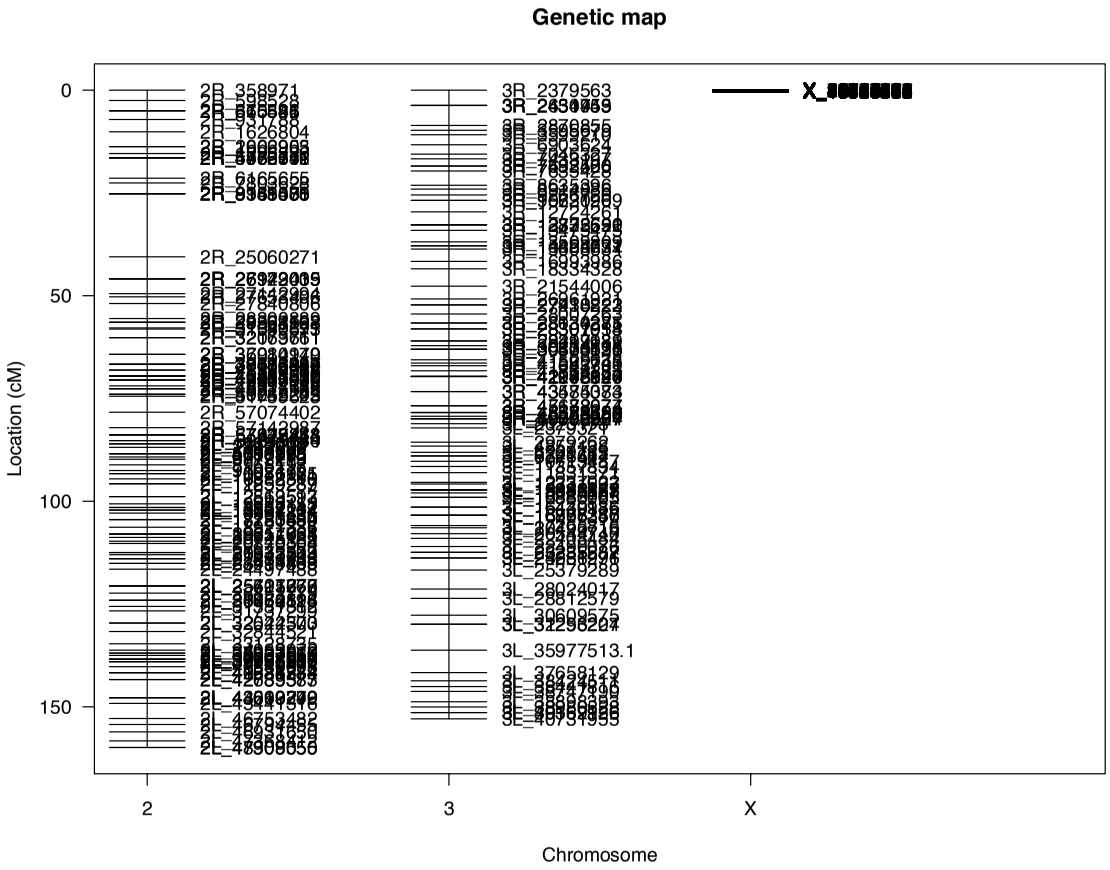


Figure 27 Genetic map of the CQxQ backross. The location of markers on chromosomes 2, 3, and X is plotted in centimorgans (cM). Marker names are plotted next to their location on the map. The marker order reflects the true order in the AgamP4.4 reference genome.

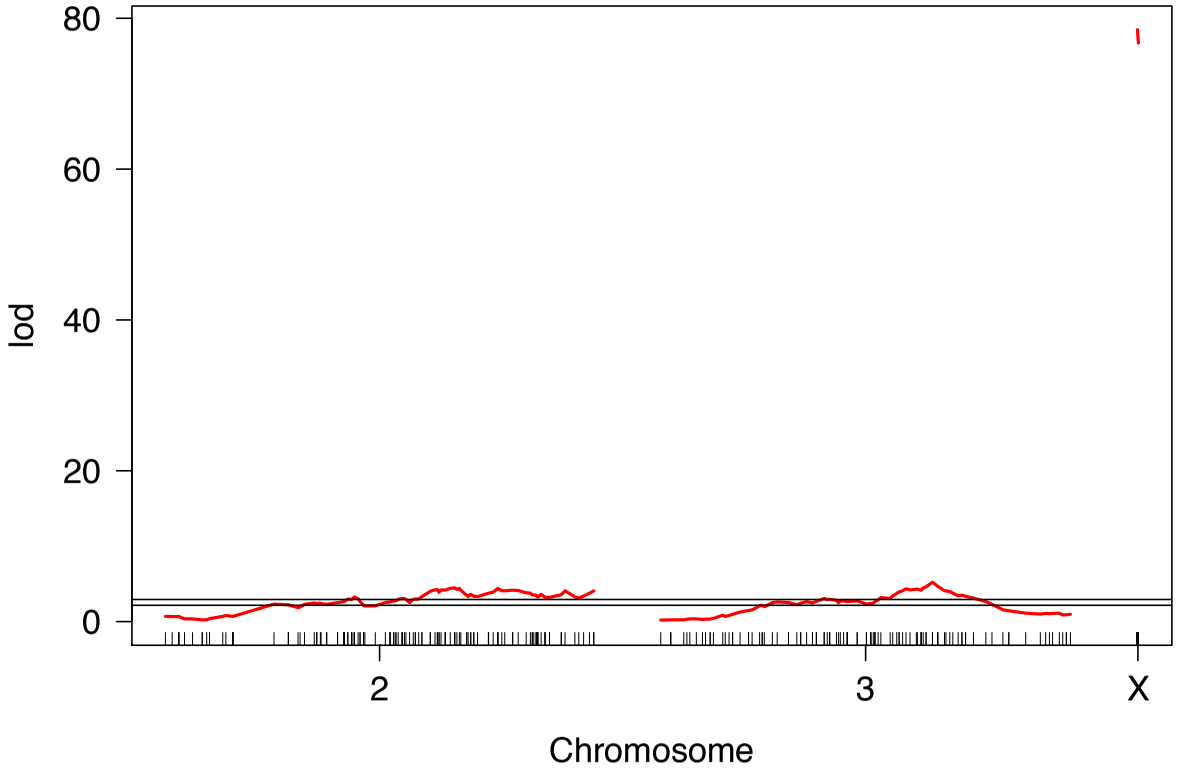


Figure 28 Genome-wide scan of logarithm of the odds (LOD) scores vs. chromosomal position for the male hybrid sterility phenotype. Solid red lines represent the results of simple interval mapping. Horizontal, black lines indicate the 5% (lower) and 1% (upper) significance threshold determined by analyzing 10,000 permutations of the simple interval mapping data.

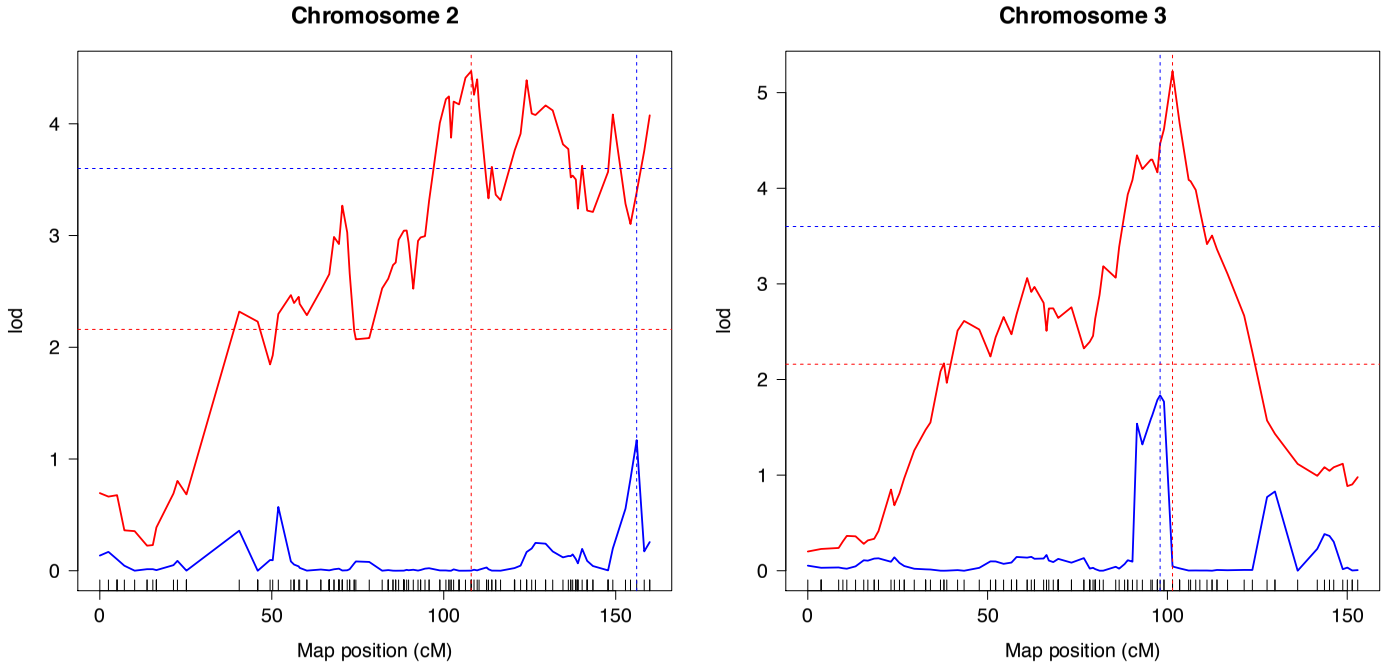


Figure 29 Genome-wide scan of logarithm of the odds (LOD) scores vs. chromosomal position for the male hybrid sterility phenotype. Solid red lines represent the results of simple interval mapping. Vertical, dotted red lines indicate the location of the marker with the highest LOD score for the SIM analysis. Horizontal, dotted red lines indicate the 5% significance threshold determined by analyzing 10,000 permutations of the SIM data. Results in blue indicate the same values resulting from the composite interval mapping analysis.

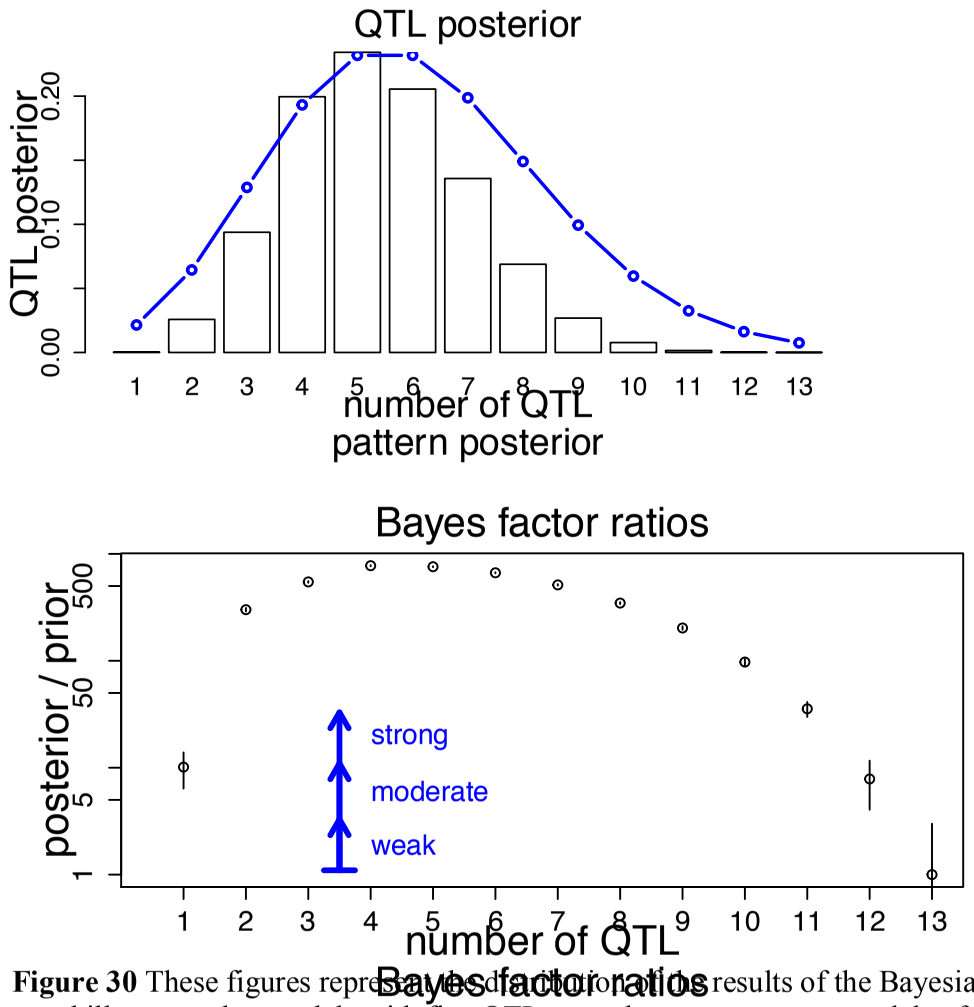


Figure 30 These figures represent the distribution of the results of the Bayesian Interval Mapping MCMC chain. The upper panel illustrates that models with five QTL were the most common models of the MCMC chain. The bottom graph shows that the Bayes factor ratios (posterior / prior) reach their asymptote between four and five QTL.

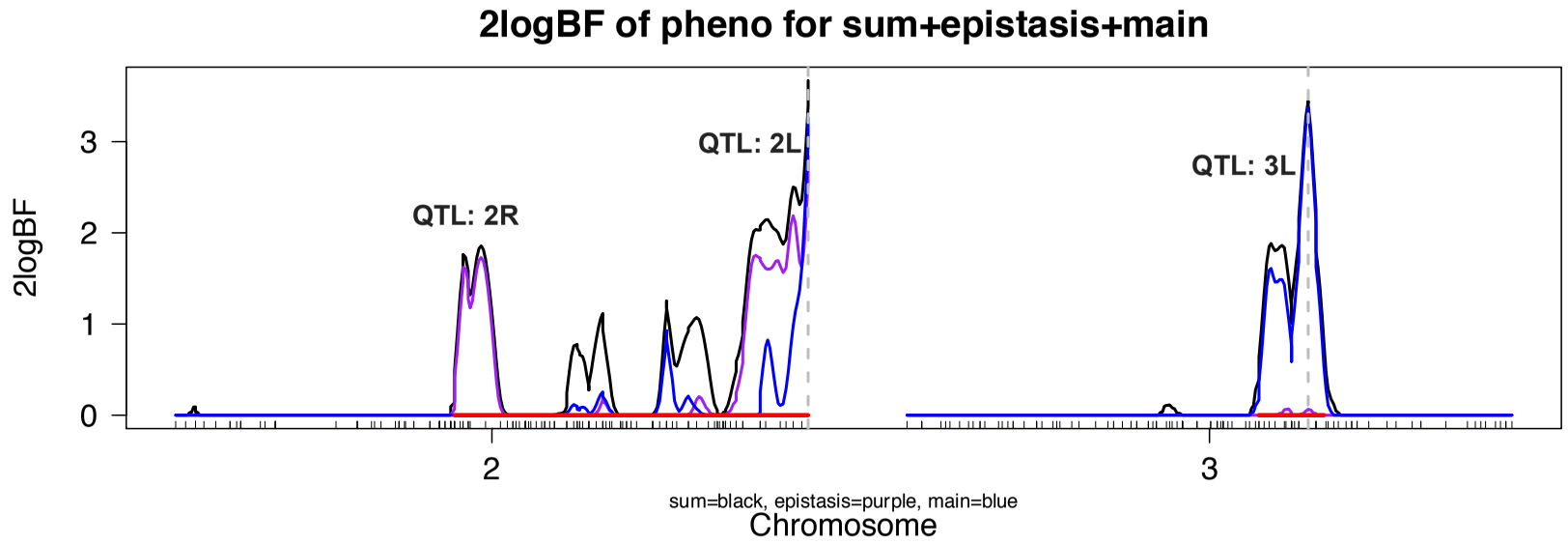


Figure 31 Genome-wide scan of logarithm 2log Bayes Factors vs. chromosomal position for the male hybrid sterility phenotype for the model with the highest posterior density (highest likelihood) in the Bayesian interval mapping analysis. Red lines indicate gene by environment interactions, which were not analyzed. Blue lines indicate main effect QTL, purple lines indicate epistatic QTL effects, and black lines indicate the sum of the two effects (main + epistasis).

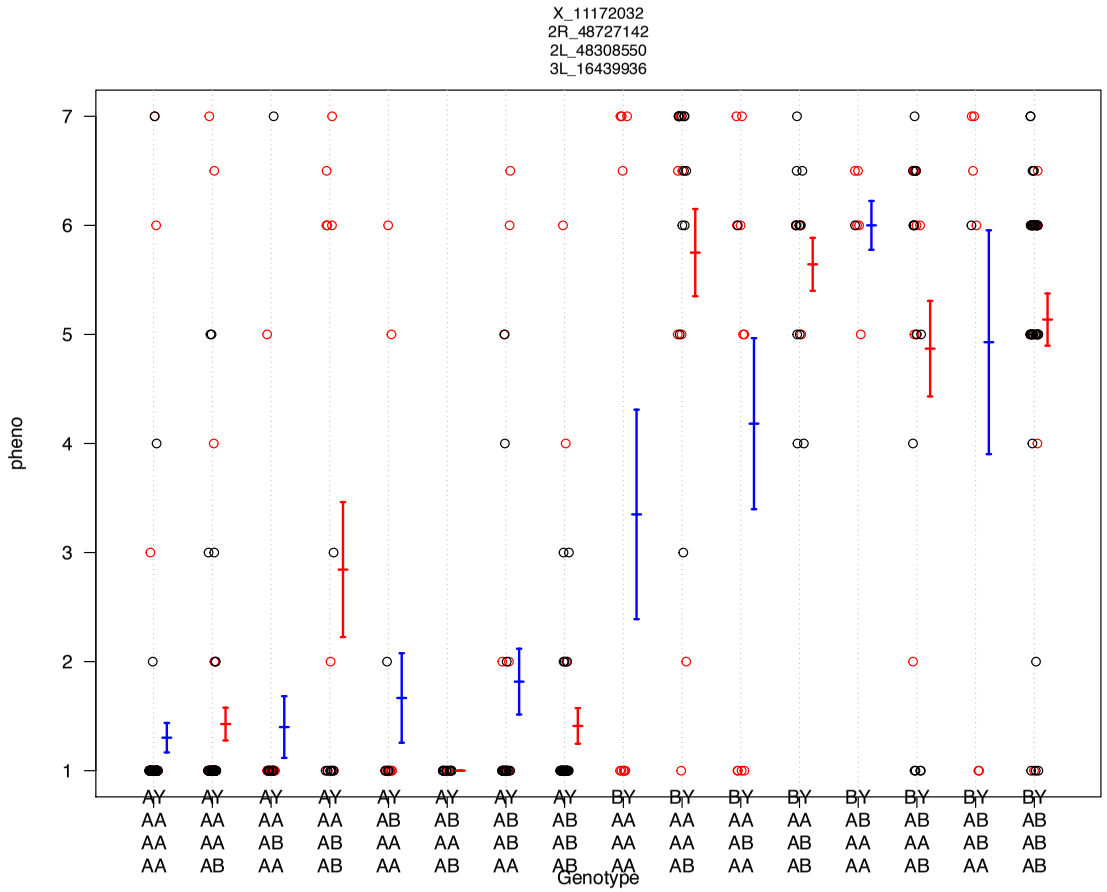


Figure 32 The effect plots the three largest QTL identified in Bayesian interval mapping (2R, 2L, and 3L) and the X chromosome. The x-axis indicates the genotype of loci in question, which are correspondingly ordered at the top of the plot. Black dots indicate observed genotypes, while red plots are imputed. Blue and red bars do not differ. The effect of the genotypes on the sterility phenotype is plotted on the y-axis.

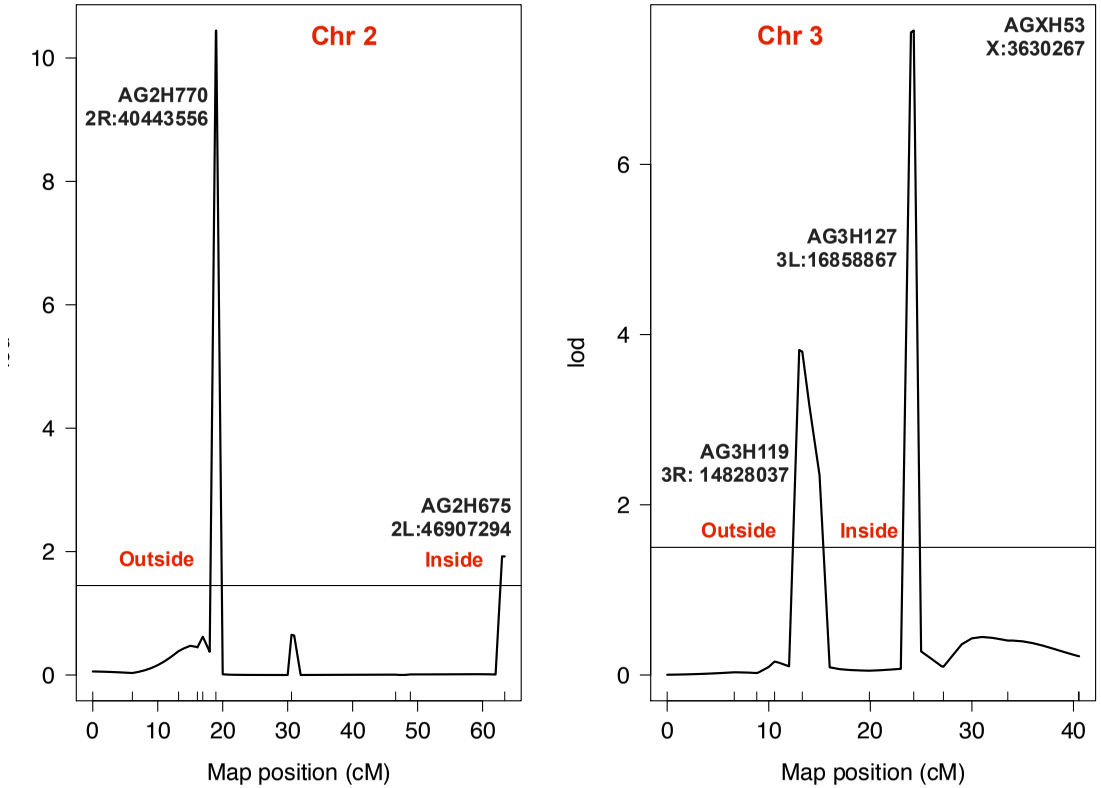


Figure 33 Genome-wide scan of logarithm of the odds (LOD) scores vs. chromosomal position for the male hybrid sterility phenotype using the QTL mapping data from the (*An. coluzzi* x *An. arabiensis*) x *An. arabiensis* hybrid cross (Slotman *et al.*, 2004. Thes are the results of composite interval mapping. Horizontal lines indicate 5% significance thresholds. Significant QTL are annotated as falling inside or outside CQxQ male sterility QTL. The microsatellite marker with the highest LOD score association with each QTL, and the genomic location of that marker, are indicated.

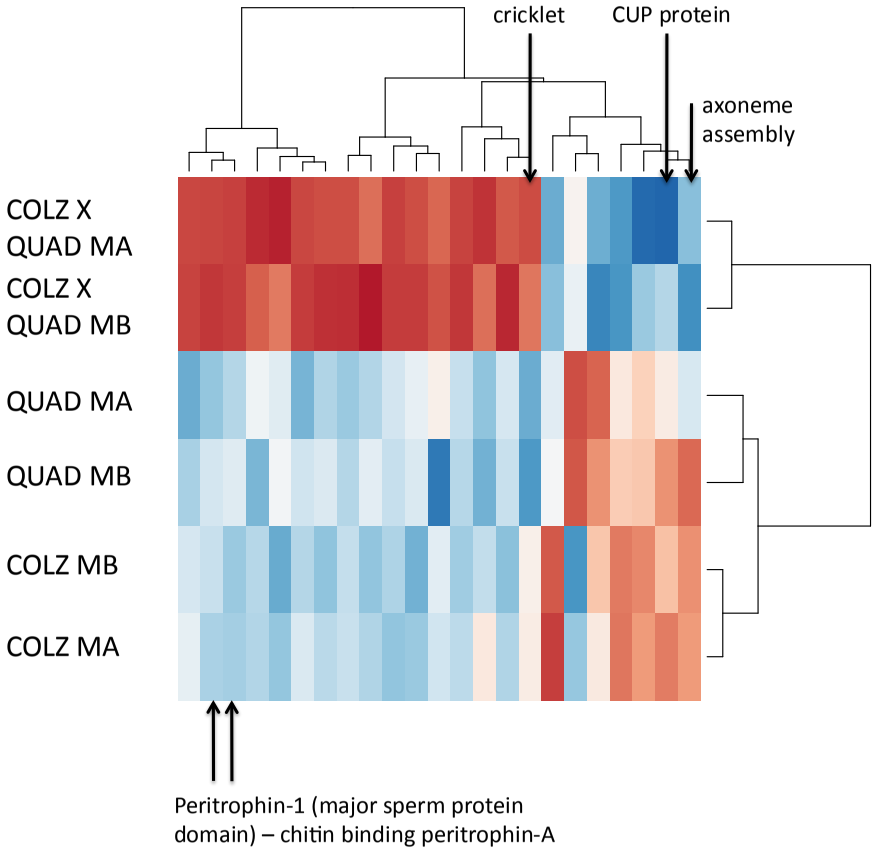


Figure 34 An expression heat map of twenty-three sex-biased in QUAD genes within sterility QTL that are significantly mis-expressed in COLZ x QUAD hybrid males in comparison to both QUAD and COLZ parental strains. Red colors show relative up-regulation, and cool colors show relative down-regulation. Samples are clustered according to expression similarity on the y-axis (right), and genes are clustered according to expression similarity along the x-axis (top). Genes with functions related to reproduction are annotated.